

# Descriptive Inference for Text

## Sophistication, Style and Statistical Properties

Arthur Spirling

New York University

August 3, 2018

# 1. The Temptation of Unsupervised Learning

# Recall

0

# Recall

Unsupervised techniques:

# Recall

Unsupervised techniques: learning  
(hidden or latent) structure in  
unlabeled data.

e.g. PCA of legislators's votes:

# Recall

Unsupervised techniques: learning  
(hidden or latent) structure in  
unlabeled data.

e.g. PCA of legislators's votes: want to see  
how they are organized—

# Recall

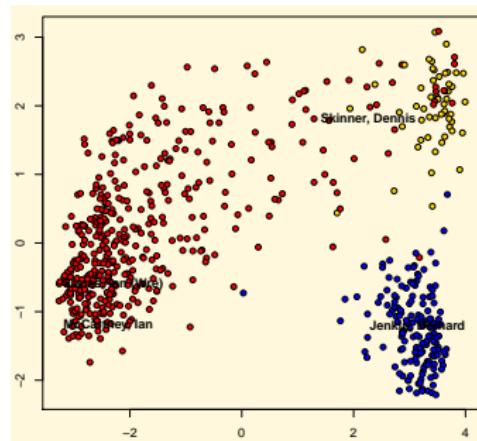
Unsupervised techniques: learning  
(hidden or latent) structure in  
unlabeled data.

- e.g. PCA of legislators's votes: want to see how they are organized—by party? by ideology? by race?

# Recall

Unsupervised techniques: learning  
(hidden or latent) structure in  
unlabeled data.

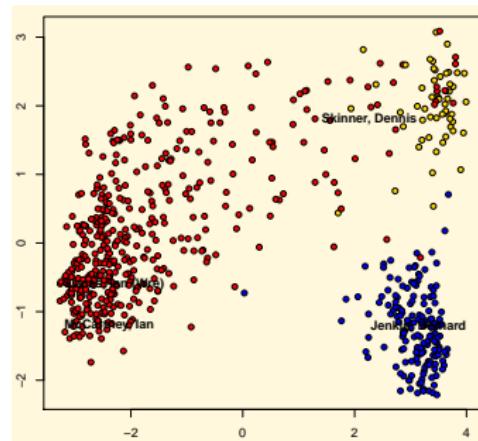
e.g. PCA of legislators's votes: want to see  
how they are organized—by party? by  
ideology? by race?



# Recall

Unsupervised techniques: learning (hidden or latent) structure in unlabeled data.

e.g. PCA of legislators's votes: want to see how they are organized—by party? by ideology? by race?

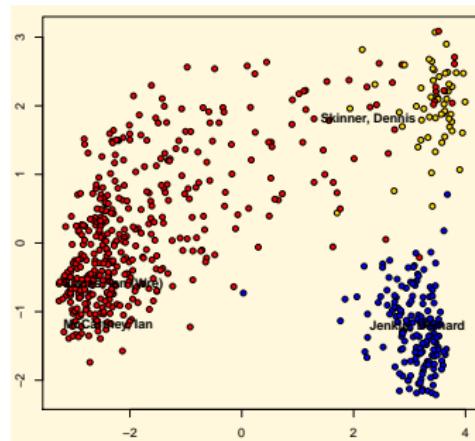


Supervised techniques:

# Recall

Unsupervised techniques: learning (hidden or latent) structure in unlabeled data.

e.g. PCA of legislators's votes: want to see how they are organized—by party? by ideology? by race?

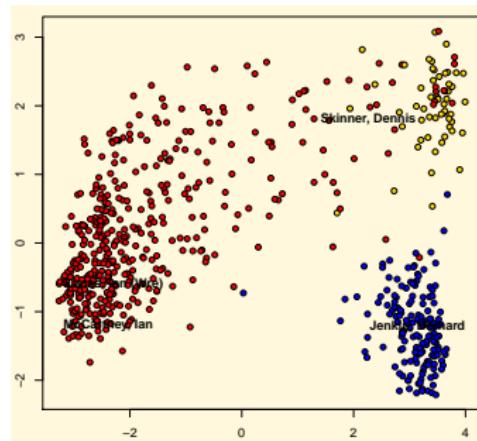


Supervised techniques: learning relationship between inputs and a labeled set of outputs.

# Recall

Unsupervised techniques: learning (hidden or latent) structure in unlabeled data.

e.g. PCA of legislators's votes: want to see how they are organized—by party? by ideology? by race?



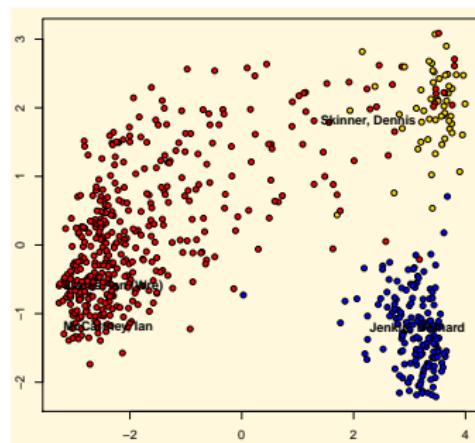
Supervised techniques: learning relationship between inputs and a labeled set of outputs.

e.g. opinion mining:

# Recall

Unsupervised techniques: learning (hidden or latent) structure in unlabeled data.

e.g. PCA of legislators's votes: want to see how they are organized—by party? by ideology? by race?



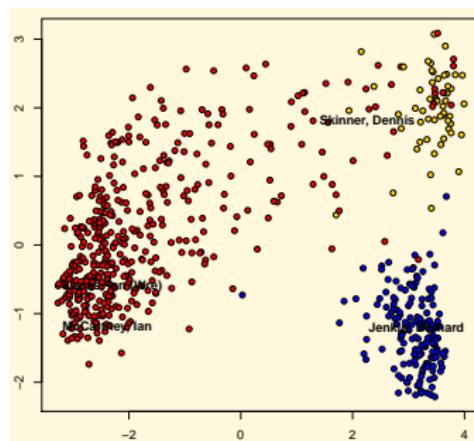
Supervised techniques: learning relationship between inputs and a labeled set of outputs.

e.g. opinion mining: what makes a critic like or dislike a movie ( $y \in \{0, 1\}$ )?

# Recall

Unsupervised techniques: learning (hidden or latent) structure in unlabeled data.

e.g. PCA of legislators's votes: want to see how they are organized—by party? by ideology? by race?



Supervised techniques: learning relationship between inputs and a labeled set of outputs.

e.g. opinion mining: what makes a critic like or dislike a movie ( $y \in \{0, 1\}$ )?

**CRITIC REVIEWS FOR STAR WARS: EPISODE VII - THE FORCE AWAKENS**

All Critics (313) | Top Critics (48) | My Critics | Fresh (293) | Rotten (20)

The new movie, as an act of pure storytelling, streams by with fluency and zip.  
[Full Review...](#) | December 21, 2015

Anthony Lane  
New Yorker  
★ Top Critic

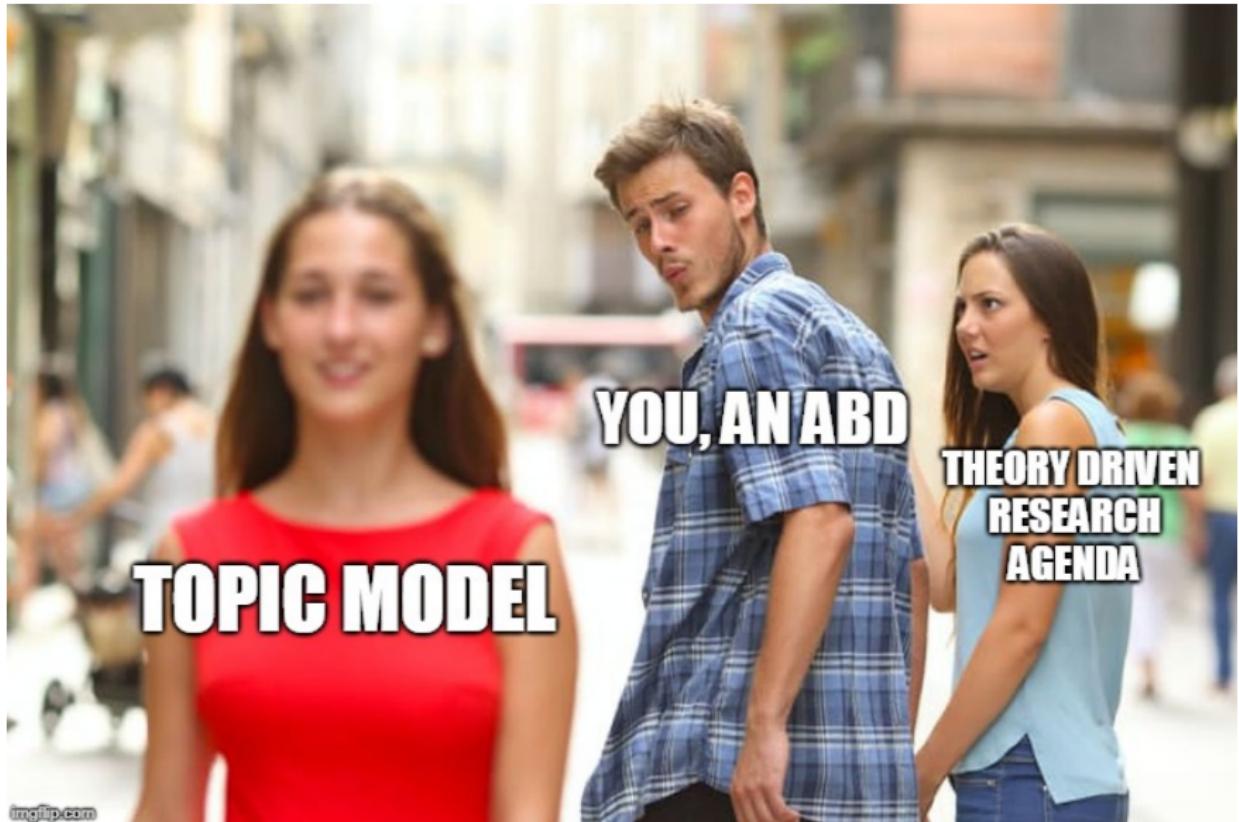
While Star Wars: The Force Awakens gets temporarily bogged down taking us back to the world that we left in 1983, it introduces us to the new and exciting torch-bearers of the franchise.  
[Full Review...](#) | December 30, 2015

Blake Howard  
Graffiti With Punctuation

At the end The Force Awakens looks more like a nostalgic film that will work as a transition to the new Star Wars' age. [Full Review in Spanish]  
[Full Review...](#) | December 29, 2015

Salvador Franco Reyes

This film is a well-planned product that balances nostalgia with the capacity to attract new generations into the Star Wars universe. [Full Review in Spanish]  
[Full Review...](#) | December 29, 2015



## Not a panacea...

'Discovery problem'—philosophically difficult to know what's been uncovered (+ hard to publish!)

## Not a panacea...

'Discovery problem'—philosophically difficult to know what's been uncovered (+ hard to publish!)

Theory wrt feature selection choices often weak

## Not a panacea...

'Discovery problem'—philosophically difficult to know what's been uncovered (+ hard to publish!)

Theory wrt feature selection choices often weak

Technical difficulties: multiple possible partitions/modes for non-determinant algorithms (even outside of topic models)

## Not a panacea...

'Discovery problem'—philosophically difficult to know what's been uncovered (+ hard to publish!)

Theory wrt feature selection choices often weak

Technical difficulties: multiple possible partitions/modes for non-determinant algorithms (even outside of topic models)

Hard to match statistical model/algorithm (black box?) to model of human behavior

## Not a panacea...

'Discovery problem'—philosophically difficult to know what's been uncovered (+ hard to publish!)

Theory wrt feature selection choices often weak

Technical difficulties: multiple possible partitions/modes for non-determinant algorithms (even outside of topic models)

Hard to match statistical model/algorithm (black box?) to model of human behavior

Statistical properties rarely discussed in measurement sense (today)

## 2. Sophistication

# Lexical Diversity

## Lexical Diversity

Recall that the elementary components of a text are called **tokens**.

## Lexical Diversity

Recall that the elementary components of a text are called **tokens**. These are generally **words**, but they may also include numbers, sums of money, etc.

## Lexical Diversity

Recall that the elementary components of a text are called **tokens**. These are generally **words**, but they may also include numbers, sums of money, etc.

The **types** in a document are the set of **unique** tokens.

## Lexical Diversity

Recall that the elementary components of a text are called **tokens**. These are generally **words**, but they may also include numbers, sums of money, etc.

The **types** in a document are the set of **unique** tokens.

thus we typically have many more tokens than types,

## Lexical Diversity

Recall that the elementary components of a text are called **tokens**. These are generally **words**, but they may also include numbers, sums of money, etc.

The **types** in a document are the set of **unique** tokens.

thus we typically have many more tokens than types, because authors **repeat** tokens.

# Lexical Diversity

Recall that the elementary components of a text are called **tokens**. These are generally **words**, but they may also include numbers, sums of money, etc.

The **types** in a document are the set of **unique** tokens.

thus we typically have many more tokens than types, because authors **repeat** tokens.

TTR we can use the **type-to-token ratio** as a measure of lexical diversity.

This is:

# Lexical Diversity

Recall that the elementary components of a text are called **tokens**. These are generally **words**, but they may also include numbers, sums of money, etc.

The **types** in a document are the set of **unique** tokens.

thus we typically have many more tokens than types, because authors **repeat** tokens.

TTR we can use the **type-to-token ratio** as a measure of lexical diversity.

This is:

$$TTR = \frac{\text{total types}}{\text{total tokens}}$$

# Lexical Diversity

Recall that the elementary components of a text are called **tokens**. These are generally **words**, but they may also include numbers, sums of money, etc.

The **types** in a document are the set of **unique** tokens.

thus we typically have many more tokens than types, because authors **repeat** tokens.

TTR we can use the **type-to-token ratio** as a measure of lexical diversity.

This is:

$$TTR = \frac{\text{total types}}{\text{total tokens}}$$

e.g. authors with limited vocabularies will have a **low** lexical diversity.

# Tabloid vs Broadsheet

# Tabloid vs Broadsheet

SEARCH

**NEW YORK POST**

NEWS

**Iraqi troops retake key government complex in Ramadi**

By Associated Press December 28, 2015 | 6:34am | Updated



Members of Iraq's elite counter-terrorism service secure a neighborhood in the city of Ramadi.

Photo: Getty Images.

**MORE ON:**  
**ISIS**

BAGHDAD — Iraqi military forces on Monday retook a strategic government complex in the city of

# Tabloid vs Broadsheet

SEARCH

**NEW YORK POST**

NEWS

**Iraqi troops retake key government complex in Ramadi**

By Associated Press December 28, 2015 | 6:34am | Updated



Members of Iraq's elite counter-terrorism service secure a neighborhood in the city of Ramadi.

Photo: Getty Images.

MORE ON:  
**ISIS**

BAGHDAD — Iraqi military forces on Monday retook a strategic government complex in the city of

$$TTR = \frac{250}{491} = 0.51$$

# Tabloid vs Broadsheet

**NEW YORK POST**

**NEWS**

## Iraqi troops retake key government complex in Ramadi

By Associated Press December 28, 2015 | 6:34am | Updated



Members of Iraq's elite counter-terrorism service secure a neighborhood in the city of Ramadi.

Photo: Getty Images.

**MORE ON: ISIS**

BAGHDAD — Iraqi military forces on Monday retook a strategic government complex in the city of

**The New York Times**

**MIDDLE EAST**

## Iraqi Forces Retake Center of Ramadi From ISIS

By FALIH HASSAN and SEWELL CHAN DEC 28, 2015



Iraqi soldiers at the Anbar police headquarters in Ramadi, Iraq, on Monday, after seizing a government complex from the Islamic State. Ahmad Al-Rubaye/Agence France-Presse — Getty Images

**Email**

**Share**

BAGHDAD — Iraqi forces said on Monday they had seized a strategic government complex in the western city of Ramadi from the Islamic State after a fierce [weeklong battle](#), putting them on the verge of a crucial victory following a brutal seven-month occupation of the city by the extremist group.

$$TTR = \frac{250}{491} = 0.51$$

# Tabloid vs Broadsheet

**NEW YORK POST**

**NEWS**

## Iraqi troops retake key government complex in Ramadi

By Associated Press December 28, 2015 | 6:34am | Updated



Members of Iraq's elite counter-terrorism service secure a neighborhood in the city of Ramadi.

Photo: Getty Images.

**MORE ON: ISIS**

BAGHDAD — Iraqi military forces on Monday retook a strategic government complex in the city of

**The New York Times**

**MIDDLE EAST**

## Iraqi Forces Retake Center of Ramadi From ISIS

By FALIH HASSAN and SEWELL CHAN DEC 28, 2015



Iraqi soldiers at the Anbar police headquarters in Ramadi, Iraq, on Monday, after seizing a government complex from the Islamic State. Ahmad Al-Rubaye/Agence France-Presse — Getty Images

BAGHDAD — Iraqi forces said on Monday they had seized a strategic government complex in the western city of Ramadi from the Islamic State after a fierce **weeklong battle**, putting them on the verge of a crucial victory following a brutal seven-month occupation of the city by the extremist group.

$$TTR = \frac{250}{491} = 0.51$$

$$TTR = \frac{428}{978} = 0.43$$

Hmm...

Hmm...

Unexpected, and mostly product of different text [lengths](#):

Hmm...

Unexpected, and mostly product of different text [lengths](#): shorter texts tend to have fewer repetitions (of e.g. common words).

Hmm...

Unexpected, and mostly product of different text [lengths](#): shorter texts tend to have fewer repetitions (of e.g. common words).

but also case that longer documents cover more topics which presumably adds to richness (?)

Hmm...

Unexpected, and mostly product of different text [lengths](#): shorter texts tend to have fewer repetitions (of e.g. common words).

[but](#) also case that longer documents cover more topics which presumably adds to richness (?)

[so](#) make denominator non-linear:

Hmm...

Unexpected, and mostly product of different text [lengths](#): shorter texts tend to have fewer repetitions (of e.g. common words).

but also case that longer documents cover more topics which presumably adds to richness (?)

so make denominator non-linear:

1954 Guiraud index of lexical richness :

$$R = \frac{\text{total types}}{\sqrt{\text{total tokens}}}$$

Hmm...

Unexpected, and mostly product of different text [lengths](#): shorter texts tend to have fewer repetitions (of e.g. common words).

but also case that longer documents cover more topics which presumably adds to richness (?)

so make denominator non-linear:

1954 Guiraud index of lexical richness :

$$R = \frac{\text{total types}}{\sqrt{\text{total tokens}}}$$

so NY Post:  $\frac{250}{\sqrt{491}} = 11.28$  ;

Hmm...

Unexpected, and mostly product of different text [lengths](#): shorter texts tend to have fewer repetitions (of e.g. common words).

but also case that longer documents cover more topics which presumably adds to richness (?)

so make denominator non-linear:

1954 Guiraud index of lexical richness :

$$R = \frac{\text{total types}}{\sqrt{\text{total tokens}}}$$

so NY Post:  $\frac{250}{\sqrt{491}} = 11.28$  ; NYT:  $\frac{428}{\sqrt{978}} = 13.68$ .

Hmm...

Unexpected, and mostly product of different text [lengths](#): shorter texts tend to have fewer repetitions (of e.g. common words).

but also case that longer documents cover more topics which presumably adds to richness (?)

so make denominator non-linear:

1954 Guiraud index of lexical richness :

$$R = \frac{\text{total types}}{\sqrt{\text{total tokens}}}$$

so NY Post:  $\frac{250}{\sqrt{491}} = 11.28$  ; NYT:  $\frac{428}{\sqrt{978}} = 13.68$ .

→ has been augmented

Hmm...

Unexpected, and mostly product of different text [lengths](#): shorter texts tend to have fewer repetitions (of e.g. common words).

but also case that longer documents cover more topics which presumably adds to richness (?)

so make denominator non-linear:

1954 Guiraud index of lexical richness :

$$R = \frac{\text{total types}}{\sqrt{\text{total tokens}}}$$

so NY Post:  $\frac{250}{\sqrt{491}} = 11.28$  ; NYT:  $\frac{428}{\sqrt{978}} = 13.68$ .

→ has been augmented—[Advanced Guiraud](#)—to exclude very common words.

# Partner Exercise

## Partner Exercise

*Restoration of national income, which shows continuing gains for the third successive year, supports the normal and logical policies under which agriculture and industry are returning to full activity. Under these policies we approach a balance of the national budget. National income increases; tax receipts, based on that income, increase without the levying of new taxes.*

*Some say my tax plan is too big. Others say its too small. I respectfully disagree.*

## Partner Exercise

*Restoration of national income, which shows continuing gains for the third successive year, supports the normal and logical policies under which agriculture and industry are returning to full activity. Under these policies we approach a balance of the national budget. National income increases; tax receipts, based on that income, increase without the levying of new taxes.*

*Some say my tax plan is too big. Others say its too small. I respectfully disagree.*

Compare these two speech segments. Which is more difficult to understand?

## Partner Exercise

*Restoration of national income, which shows continuing gains for the third successive year, supports the normal and logical policies under which agriculture and industry are returning to full activity. Under these policies we approach a balance of the national budget. National income increases; tax receipts, based on that income, increase without the levying of new taxes.*

*Some say my tax plan is too big. Others say its too small. I respectfully disagree.*

Compare these two speech segments. Which is more difficult to understand? Why: which features are important?

# Measurement of Linguistic Complexity

# Measurement of Linguistic Complexity

- over a hundred years of literature on measurement of 'readability':

# Measurement of Linguistic Complexity

- over a hundred years of literature on measurement of 'readability': general issue was assigning school texts to pupils of different ages and abilities.

# Measurement of Linguistic Complexity

- over a hundred years of literature on measurement of 'readability': general issue was assigning school texts to pupils of different ages and abilities.
- Flesch (1948) suggests *Flesch Reading Ease* statistic

# Measurement of Linguistic Complexity

- over a hundred years of literature on measurement of 'readability': general issue was assigning school texts to pupils of different ages and abilities.
- Flesch (1948) suggests *Flesch Reading Ease* statistic

## FRE

$$= 206.835 - 1.015 \left( \frac{\text{total words}}{\text{total sentences}} \right) - 84.6 \left( \frac{\text{total syllables}}{\text{total words}} \right)$$

# Measurement of Linguistic Complexity

- over a hundred years of literature on measurement of 'readability': general issue was assigning school texts to pupils of different ages and abilities.
- Flesch (1948) suggests *Flesch Reading Ease* statistic

## FRE

$$= 206.835 - 1.015 \left( \frac{\text{total words}}{\text{total sentences}} \right) - 84.6 \left( \frac{\text{total syllables}}{\text{total words}} \right)$$

based on  $\hat{\beta}$ s from linear model where  $y$  = average grade level of school children who could correctly answer at least 75% of mc qs on texts.

# Measurement of Linguistic Complexity

- over a hundred years of literature on measurement of 'readability': general issue was assigning school texts to pupils of different ages and abilities.
- Flesch (1948) suggests *Flesch Reading Ease* statistic

## FRE

$$= 206.835 - 1.015 \left( \frac{\text{total words}}{\text{total sentences}} \right) - 84.6 \left( \frac{\text{total syllables}}{\text{total words}} \right)$$

based on  $\hat{\beta}$ s from linear model where  $y$  = average grade level of school children who could correctly answer at least 75% of mc qs on texts. Scaled s.t. a document with score of 100 could be understood by fourth grader (9yo).

# Measurement of Linguistic Complexity

- over a hundred years of literature on measurement of 'readability': general issue was assigning school texts to pupils of different ages and abilities.
- Flesch (1948) suggests *Flesch Reading Ease* statistic

## FRE

$$= 206.835 - 1.015 \left( \frac{\text{total words}}{\text{total sentences}} \right) - 84.6 \left( \frac{\text{total syllables}}{\text{total words}} \right)$$

based on  $\hat{\beta}$ s from linear model where  $y$  = average grade level of school children who could correctly answer at least 75% of mc qs on texts. Scaled s.t. a document with score of 100 could be understood by fourth grader (9yo).

- Kincaid et al later translate to US School *grade level* that would be (on average) required to comprehend text.

# Readability Guidelines

# Readability Guidelines

in practice,

# Readability Guidelines

in practice, estimated FRE can be outside [0, 100].

## Readability Guidelines

in practice, estimated FRE can be outside [0, 100].

However...

# Readability Guidelines

in practice, estimated FRE can be outside [0, 100].

However...

Score	Education	Description	Clve % US popn
0–30	college graduates	very difficult	28
31–50		difficult	72
51–60		fairly difficult	85
61–70	9th grade	standard	—
71–80		fairly easy	—
81–90		easy	—
91–100	4th grade	very easy	—

# Examples

## Examples

Score	Text
-800	Molly Bloom's (3.6K) Soliloquy, <i>Ulysses</i>

## Examples

Score	Text
-800	Molly Bloom's (3.6K) Soliloquy, <i>Ulysses</i>
33	mean political science article; judicial opinion

## Examples

Score	Text
-800	Molly Bloom's (3.6K) Soliloquy, <i>Ulysses</i>
33	mean political science article; judicial opinion
37	Spirling
45	life insurance requirement (FL)

## Examples

Score	Text
-800	Molly Bloom's (3.6K) Soliloquy, <i>Ulysses</i>
33	mean political science article; judicial opinion
37	<b>Spirling</b>
45	life insurance requirement (FL)
48	<i>New York times</i>
65	<i>Reader's Digest</i>
67	<i>Al Qaeda</i> press release

## Examples

Score	Text
-800	Molly Bloom's (3.6K) Soliloquy, <i>Ulysses</i>
33	mean political science article; judicial opinion
37	<b>Spirling</b>
45	life insurance requirement (FL)
48	<i>New York times</i>
65	<i>Reader's Digest</i>
67	Al Qaeda press release
77	Dickens' works
80	children's books: e.g. <i>Wind in the Willows</i>

## Examples

Score	Text
-800	Molly Bloom's (3.6K) Soliloquy, <i>Ulysses</i>
33	mean political science article; judicial opinion
37	<i>Spirling</i>
45	life insurance requirement (FL)
48	<i>New York times</i>
65	<i>Reader's Digest</i>
67	Al Qaeda press release
77	Dickens' works
80	children's books: e.g. <i>Wind in the Willows</i>
90	death row inmate last statements (TX)
100	this entry right here.

# Notes

0

# Notes

Flesch scoring only uses **syllable** information:

# Notes

Flesch scoring only uses syllable information: no input from rarity or unfamiliarity of word.

## Notes

Flesch scoring only uses syllable information: no input from rarity or unfamiliarity of word.

e.g. "Indeed, the shoemaker was frightened" would score similarly to

## Notes

Flesch scoring only uses syllable information: no input from rarity or unfamiliarity of word.

e.g. "Indeed, the shoemaker was frightened" would score similarly to "Forsooth, the cordwainer was afeared"

## Notes

Flesch scoring only uses syllable information: no input from rarity or unfamiliarity of word.

e.g. "Indeed, the shoemaker was frightened" would score similarly to "Forsooth, the cordwainer was afeared"

Widely used because it 'works',

## Notes

Flesch scoring only uses syllable information: no input from rarity or unfamiliarity of word.

e.g. "Indeed, the shoemaker was frightened" would score similarly to "Forsooth, the cordwainer was afeared"

Widely used because it 'works', not because it is justified from first principles

## Notes

Flesch scoring only uses syllable information: no input from rarity or unfamiliarity of word.

e.g. "Indeed, the shoemaker was frightened" would score similarly to "Forsooth, the cordwainer was afeared"

Widely used because it 'works', not because it is justified from first principles

One of many such indices:

## Notes

Flesch scoring only uses syllable information: no input from rarity or unfamiliarity of word.

e.g. "Indeed, the shoemaker was frightened" would score similarly to "Forsooth, the cordwainer was afeared"

Widely used because it 'works', not because it is justified from first principles

One of many such indices: Gunning-Fog,

## Notes

Flesch scoring only uses syllable information: no input from rarity or unfamiliarity of word.

e.g. "Indeed, the shoemaker was frightened" would score similarly to "Forsooth, the cordwainer was afeared"

Widely used because it 'works', not because it is justified from first principles

One of many such indices: Gunning-Fog, Dale-Chall,

## Notes

Flesch scoring only uses [syllable](#) information: no input from rarity or [unfamiliarity](#) of word.

e.g. “Indeed, the shoemaker was frightened” would score similarly to “Forsooth, the cordwainer was afeared”

Widely used because it ‘works’, not because it is justified from [first principles](#)

One of [many](#) such indices: Gunning-Fog, [Dale-Chall](#), Automated Readability Index,

## Notes

Flesch scoring only uses syllable information: no input from rarity or unfamiliarity of word.

e.g. "Indeed, the shoemaker was frightened" would score similarly to "Forsooth, the cordwainer was afeared"

Widely used because it 'works', not because it is justified from first principles

One of many such indices: Gunning-Fog, Dale-Chall, Automated Readability Index, SMOG.

## Notes

Flesch scoring only uses syllable information: no input from rarity or unfamiliarity of word.

e.g. "Indeed, the shoemaker was frightened" would score similarly to "Forsooth, the cordwainer was afeared"

Widely used because it 'works', not because it is justified from first principles

One of many such indices: Gunning-Fog, Dale-Chall, Automated Readability Index, SMOG. Typically highly correlated (at text level).

## Notes

Flesch scoring only uses syllable information: no input from rarity or unfamiliarity of word.

e.g. "Indeed, the shoemaker was frightened" would score similarly to "Forsooth, the cordwainer was afeared"

Widely used because it 'works', not because it is justified from first principles

One of many such indices: Gunning-Fog, Dale-Chall, Automated Readability Index, SMOG. Typically highly correlated (at text level).

Surprisingly little effort to describe statistical behavior of estimator:

## Notes

Flesch scoring only uses syllable information: no input from rarity or unfamiliarity of word.

e.g. "Indeed, the shoemaker was frightened" would score similarly to "Forsooth, the cordwainer was afeared"

Widely used because it 'works', not because it is justified from first principles

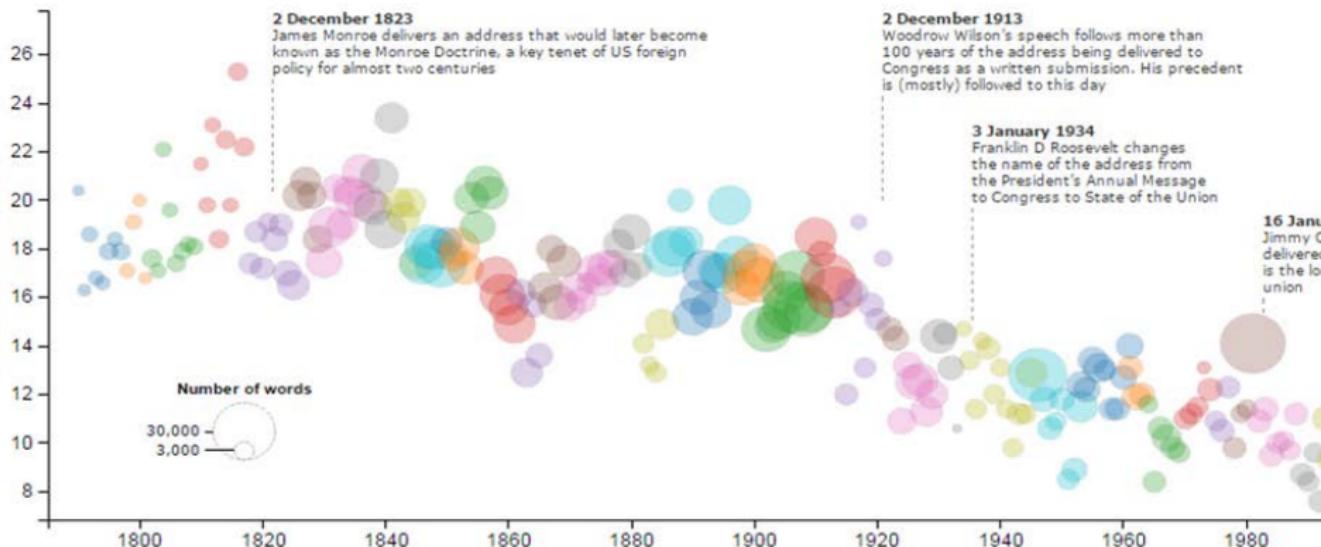
One of many such indices: Gunning-Fog, Dale-Chall, Automated Readability Index, SMOG. Typically highly correlated (at text level).

Surprisingly little effort to describe statistical behavior of estimator: sampling distribution etc.

# The state of our union is ... dumber:

## How the linguistic standard of the presidential address has declined

Using the Flesch-Kincaid readability test the Guardian has tracked the reading level of every State of the Union



# Leaders and their incentives

# Leaders and their incentives

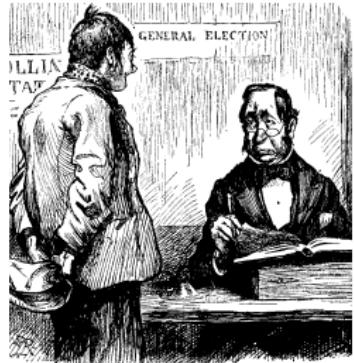
C19th Britain is notable for fast expansion of suffrage.



# Leaders and their incentives

C19th Britain is notable for fast **expansion of suffrage.**

new voters tended to be poorer and **less literate**

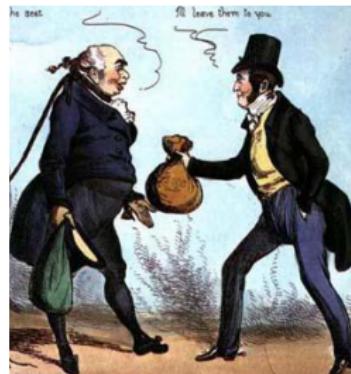


# Leaders and their incentives

C19th Britain is notable for fast expansion of suffrage.

new voters tended to be poorer and less literate

↓ local, clientelistic appeals via bribery...



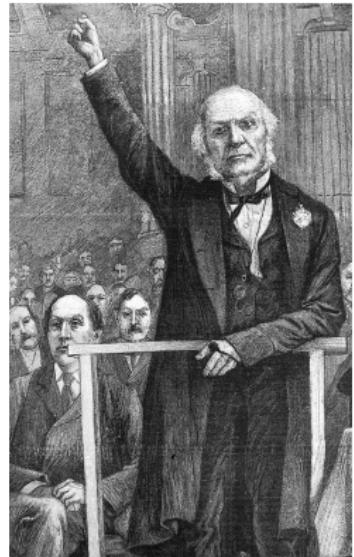
# Leaders and their incentives

C19th Britain is notable for fast **expansion of suffrage**.

new voters tended to be poorer and **less literate**

↓ local, clientalistic appeals via bribery...

↑ 'party orientated electorate', with national policies and national **leaders**



# Leaders and their incentives

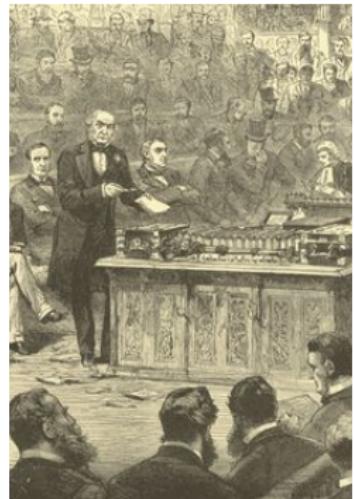
C19th Britain is notable for fast **expansion of suffrage**.

new voters tended to be poorer and **less literate**

↓ local, clientelistic appeals via bribery...

↑ 'party orientated electorate', with national policies and national **leaders**

Q how did these leaders respond to new voters?



# Leaders and their incentives

C19th Britain is notable for fast **expansion of suffrage**.

new voters tended to be poorer and **less literate**

↓ local, clientelistic appeals via bribery...

↑ 'party orientated electorate', with national policies and national **leaders**

**Q** how did these leaders **respond** to new voters?

**A** by changing nature of their speech:



# Leaders and their incentives

C19th Britain is notable for fast **expansion of suffrage**.

new voters tended to be poorer and **less literate**

↓ local, clientelistic appeals via bribery...

↑ 'party orientated electorate', with national policies and national **leaders**

**Q** how did these leaders **respond** to new voters?

**A** by changing nature of their speech: **simpler**,



# Leaders and their incentives

C19th Britain is notable for fast **expansion of suffrage**.

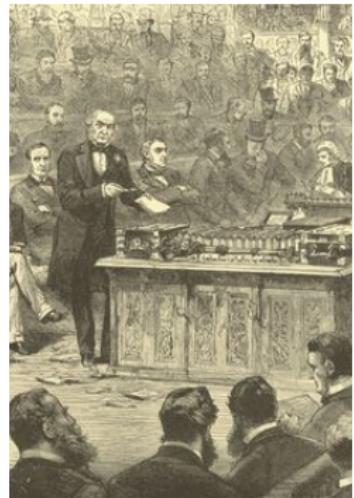
new voters tended to be poorer and **less literate**

↓ local, clientelistic appeals via bribery...

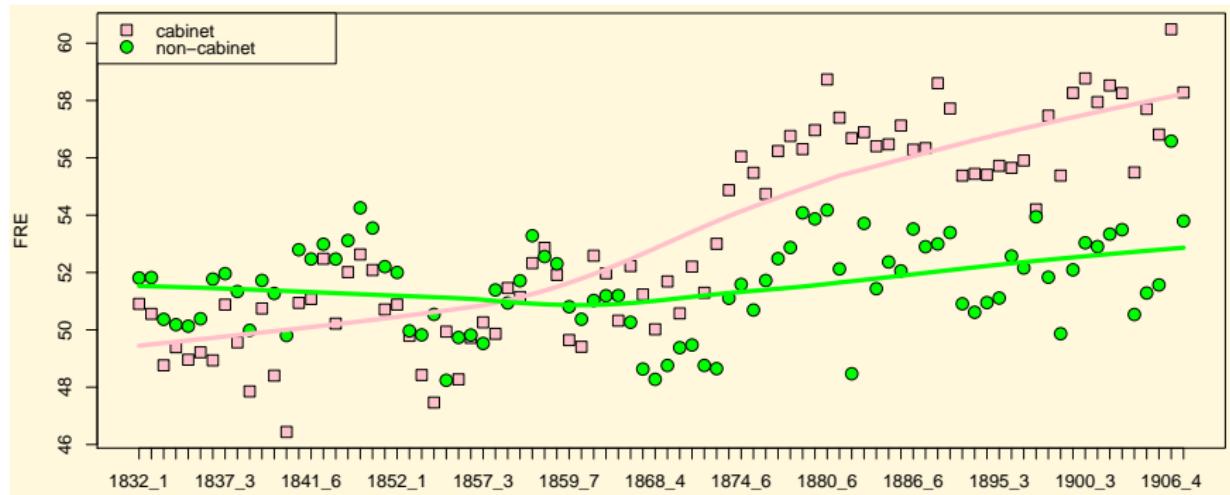
↑ 'party orientated electorate', with national policies and national **leaders**

**Q** how did these leaders **respond** to new voters?

**A** by changing nature of their speech: **simpler**, less complex expressions in parliament



# Flesch overtime plot



# Dale-Chall, 1948

# Dale-Chall, 1948

yields grade level of text sample.

## Dale-Chall, 1948

yields grade level of text sample.

DC

$$0.1579 \times (\text{PDW}) + 0.0496 \times \left( \frac{\text{total words}}{\text{total sentences}} \right)$$

## Dale-Chall, 1948

yields grade level of text sample.

DC

$$0.1579 \times (\text{PDW}) + 0.0496 \times \left( \frac{\text{total words}}{\text{total sentences}} \right)$$

where PDW is percentage of difficult words,

## Dale-Chall, 1948

yields grade level of text sample.

DC

$$0.1579 \times (\text{PDW}) + 0.0496 \times \left( \frac{\text{total words}}{\text{total sentences}} \right)$$

where PDW is percentage of difficult words,

and a 'difficult' word is one that does not appear on Dale & Chall's list of 763

## Dale-Chall, 1948

yields grade level of text sample.

DC

$$0.1579 \times (\text{PDW}) + 0.0496 \times \left( \frac{\text{total words}}{\text{total sentences}} \right)$$

where PDW is percentage of difficult words,

and a 'difficult' word is one that does not appear on Dale & Chall's list of 763 (later updated to 3000)

## Dale-Chall, 1948

yields grade level of text sample.

DC

$$0.1579 \times (\text{PDW}) + 0.0496 \times \left( \frac{\text{total words}}{\text{total sentences}} \right)$$

where PDW is percentage of difficult words,

and a 'difficult' word is one that does not appear on Dale & Chall's list of 763 (later updated to 3000) 'familiar' words.

## Dale-Chall, 1948

yields grade level of text sample.

DC

$$0.1579 \times (\text{PDW}) + 0.0496 \times \left( \frac{\text{total words}}{\text{total sentences}} \right)$$

where PDW is percentage of difficult words,

and a 'difficult' word is one that does not appear on Dale & Chall's list of 763 (later updated to 3000) 'familiar' words.

e.g. about, back, call, etc.

# Partner Exercise

# Partner Exercise



## Partner Exercise



The FRE of SOTU speeches is increasing. Why might it be difficult to make readability comparisons over time?



## Partner Exercise



The FRE of SOTU speeches is increasing. Why might it be difficult to make readability comparisons over time? (hint: when were the reading ease measures invented? are topics of speeches constant? were addresses always delivered the same way?)



## Partner Exercise



The FRE of SOTU speeches is increasing. Why might it be difficult to make readability comparisons over time? (hint: when were the reading ease measures invented? are topics of speeches constant? were addresses always delivered the same way?)



Does the nature of the decline suggest that speeches are becoming simpler for demand (i.e. voter) or supply (i.e. leader) incentive reasons?

## Partner Exercise



The FRE of SOTU speeches is increasing. Why might it be difficult to make readability comparisons over time? (hint: when were the reading ease measures invented? are topics of speeches constant? were addresses always delivered the same way?)



Does the nature of the decline suggest that speeches are becoming simpler for demand (i.e. voter) or supply (i.e. leader) incentive reasons? (hint: consider the smoothness/jaggedness of the decrease)

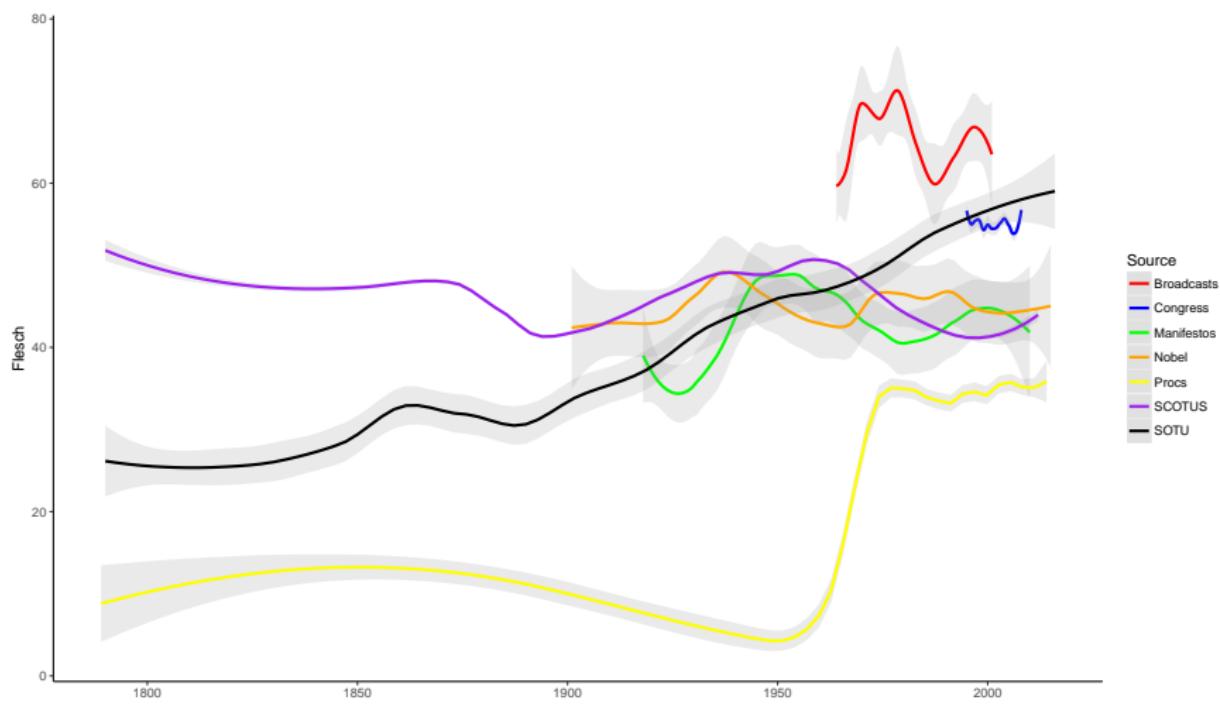
# The Great Sentence Length Shift (Benoit, Munger & Spirling)

# The Great Sentence Length Shift (Benoit, Munger & Spirling)

SOTU is simpler: is public discourse becoming 'dumbed down'? Probably not.

# The Great Sentence Length Shift (Benoit, Munger & Spirling)

SOTU is simpler: is public discourse becoming 'dumbed down'? Probably not.

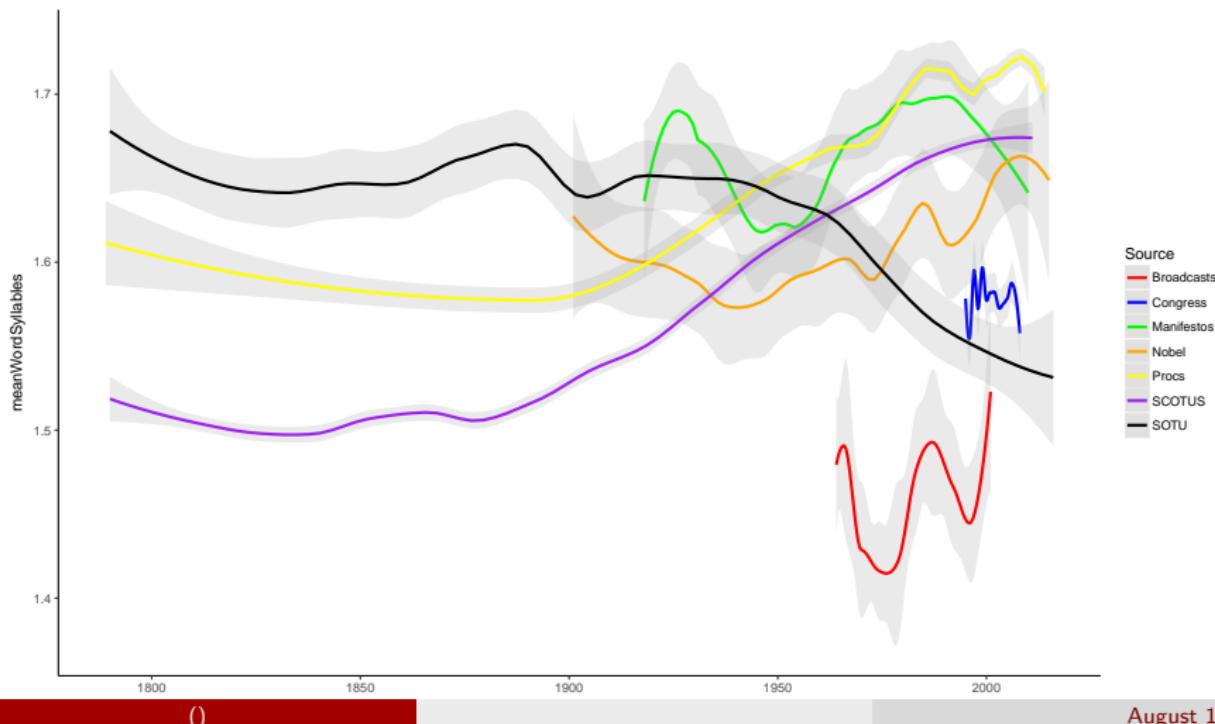


# The Great Sentence Length Shift (Benoit, Munger & Spirling)

SOTU is simpler: is public discourse becoming 'dumbed down'? Probably not.  
What's driving these patterns?

# The Great Sentence Length Shift (Benoit, Munger & Spirling)

SOTU is simpler: is public discourse becoming ‘dumbed down’? Probably not.  
What’s driving these patterns? Syllables?

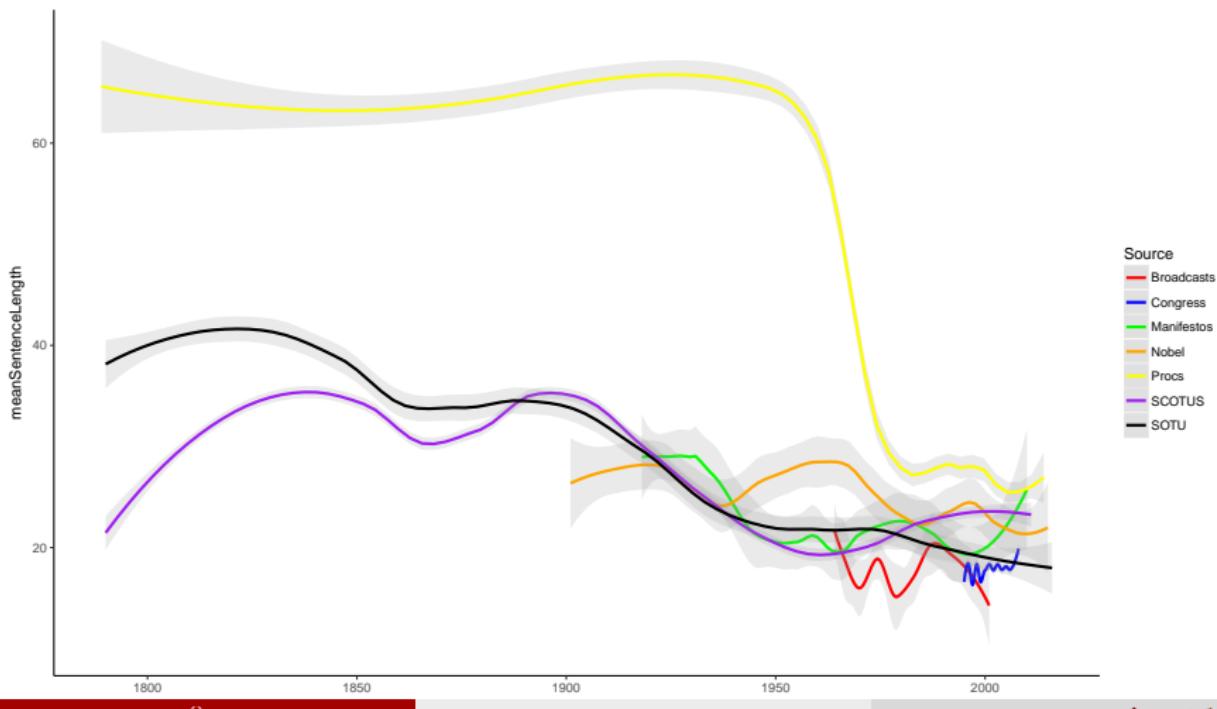


# The Great Sentence Length Shift (Benoit, Munger & Spirling)

SOTU is simpler: is public discourse becoming 'dumbed down'? Probably not.  
What's driving these patterns? Syllables? Sentence length?

# The Great Sentence Length Shift (Benoit, Munger & Spirling)

SOTU is simpler: is public discourse becoming 'dumbed down'? Probably not. What's driving these patterns? Syllables? Sentence length?



# **3. Doing Better**

Can we do better?

# Can we do better?

i.e. have I, personally, written a paper about this?

# Can we do better?

i.e. have I, personally, written a paper about this?

YES!

## Can we do better?

i.e. have I, personally, written a paper about this?

YES! BMS [crowdsource](#) thousands of snippet comparisons: ask raters which is more difficult,

## Can we do better?

i.e. have I, personally, written a paper about this?

YES! BMS **crowdsource** thousands of snippet comparisons: ask raters which is more difficult, make that a function of **covariates**.

## Can we do better?

i.e. have I, personally, written a paper about this?

YES! BMS **crowdsource** thousands of snippet comparisons: ask raters which is more difficult, make that a function of **covariates**.

Everything wrapped into well-known **statistical model** (Bradley-Terry) for pairwise comparisons: can make **probabilistic statements** about difficulty.

## Can we do better?

i.e. have I, personally, written a paper about this?

YES! BMS **crowdsource** thousands of snippet comparisons: ask raters which is more difficult, make that a function of **covariates**.

Everything wrapped into well-known **statistical model** (Bradley-Terry) for pairwise comparisons: can make **probabilistic statements** about difficulty.

Model performance not hugely better than FRE,

## Can we do better?

i.e. have I, personally, written a paper about this?

YES! BMS **crowdsource** thousands of snippet comparisons: ask raters which is more difficult, make that a function of **covariates**.

Everything wrapped into well-known **statistical model** (Bradley-Terry) for pairwise comparisons: can make **probabilistic statements** about difficulty.

Model performance not hugely better than FRE, but note we are at least able to make the comparison!

## Can we do better?

i.e. have I, personally, written a paper about this?

YES! BMS **crowdsource** thousands of snippet comparisons: ask raters which is more difficult, make that a function of **covariates**.

Everything wrapped into well-known **statistical model** (Bradley-Terry) for pairwise comparisons: can make **probabilistic statements** about difficulty.

Model performance not hugely better than FRE, but note we are at least able to make the comparison!

Can ask: what is  $\Pr(\text{Eisenhower easier than Bush})$ ? (is  $\sim 0.43$ )

## Can we do better?

i.e. have I, personally, written a paper about this?

YES! BMS **crowdsource** thousands of snippet comparisons: ask raters which is more difficult, make that a function of **covariates**.

Everything wrapped into well-known **statistical model** (Bradley-Terry) for pairwise comparisons: can make **probabilistic statements** about difficulty.

Model performance not hugely better than FRE, but note we are at least able to make the comparison!

Can ask: what is  $\Pr(\text{Eisenhower easier than Bush})$ ? (is  $\sim 0.43$ )

Key innovation:

## Can we do better?

i.e. have I, personally, written a paper about this?

YES! BMS [crowdsource](#) thousands of snippet comparisons: ask raters which is more difficult, make that a function of [covariates](#).

Everything wrapped into well-known [statistical model](#) (Bradley-Terry) for pairwise comparisons: can make [probabilistic statements](#) about difficulty.

Model performance not hugely better than FRE, but note we are at least able to make the comparison!

Can ask: what is  $\Pr(\text{Eisenhower easier than Bush})$ ? (is  $\sim 0.43$ )

Key innovation: rarity is from [google books](#) corpus,

## Can we do better?

i.e. have I, personally, written a paper about this?

YES! BMS [crowdsource](#) thousands of snippet comparisons: ask raters which is more difficult, make that a function of [covariates](#).

Everything wrapped into well-known [statistical model](#) (Bradley-Terry) for pairwise comparisons: can make [probabilistic statements](#) about difficulty.

Model performance not hugely better than FRE, but note we are at least able to make the comparison!

Can ask: what is  $\Pr(\text{Eisenhower easier than Bush})$ ? (is  $\sim 0.43$ )

Key innovation: rarity is from [google books](#) corpus, and fitted to local decade and domain (adults) that you care about.

# Details on Rarity

## Details on Rarity

[Google Books Corpus](#) contains frequency counts for  $\sim$  155 billion words, 1550–2008.

## Details on Rarity

[Google Books Corpus](#) contains frequency counts for  $\sim$  155 billion words, 1550–2008.

We calculate a frequency of a word (smoothed into decades) as the rate of use of that word *relative* to the (which is fairly constant).

## Details on Rarity

[Google Books Corpus](#) contains frequency counts for  $\sim 155$  billion words, 1550–2008.

We calculate a frequency of a word (smoothed into decades) as the rate of use of that word *relative* to the (which is fairly constant).

We had to do a [a lot](#) of cleaning: removing typos/non-words etc.

## Details on Rarity

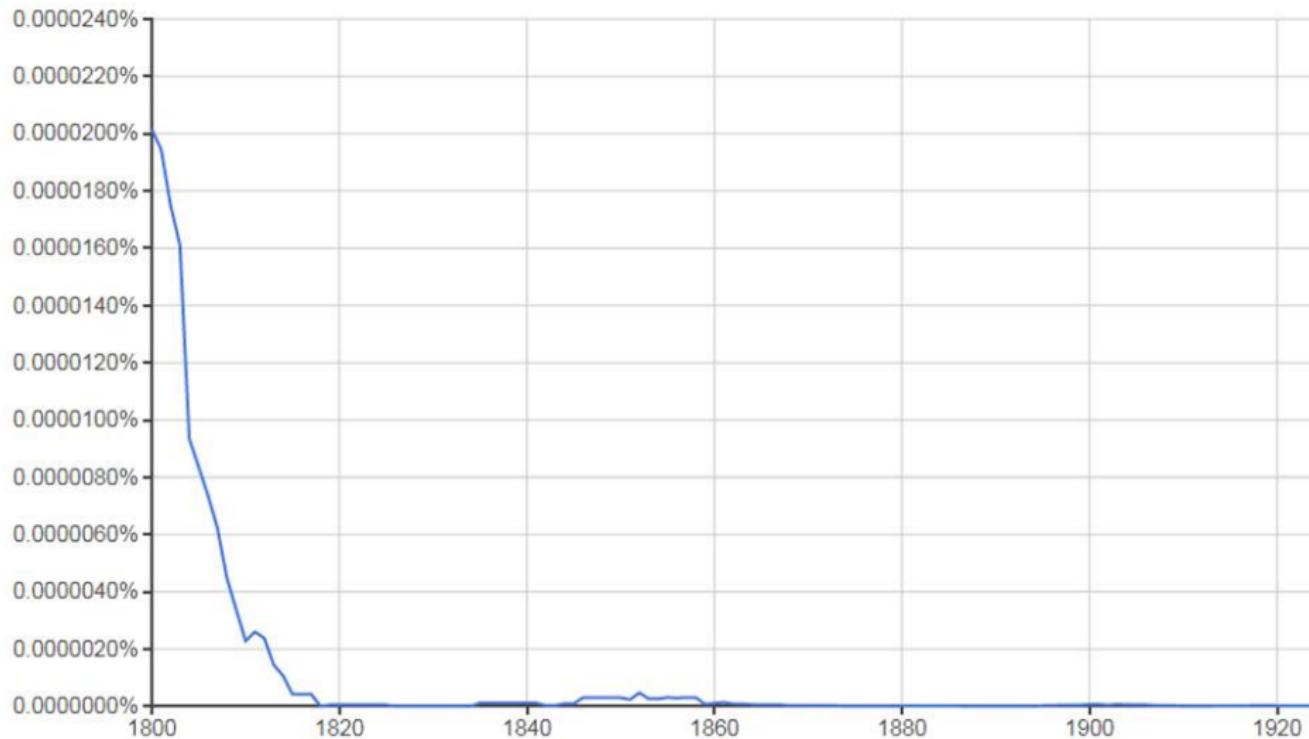
Google Books Corpus contains frequency counts for  $\sim 155$  billion words, 1550–2008.

We calculate a frequency of a word (smoothed into decades) as the rate of use of that word *relative* to the (which is fairly constant).

We had to do a *a lot* of cleaning: removing typos/non-words etc.

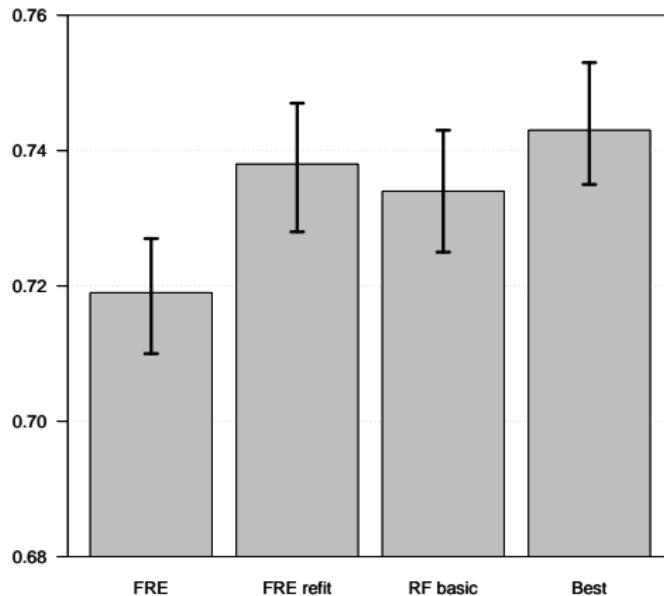
Two variables: *minimum* rarity of word in a snippet (highest when rarest word in snippet is common in corpus); *mean* rarity (lowest when average rarity is low—i.e. words in snippet are common in corpus)

## Cleaning: On the incidence of ftupid



# Performance Compared

# Performance Compared



Our best model is preferred, and offers a 'real' improvement over FRE.

# Why We Beat FRE

# Why We Beat FRE



The first **cession** was made by the State of New York, and the largest, which in area exceeded all the others, by the State of Virginia.

# Why We Beat FRE



The first **cession** was made by the State of New York, and the largest, which in area exceeded all the others, by the State of Virginia.

I speak to you not just as a President, but as a father, when I say that responsibility for our children's education must begin at home.

# Why We Beat FRE



The first **cession** was made by the State of New York, and the largest, which in area exceeded all the others, by the State of Virginia.

I speak to you not just as a President, but as a father, when I say that responsibility for our children's education must begin at home.

Cleveland wins on FRE, but Obama wins in our model (penalizing for rarity).

# Paper and Software

# Paper and Software

[Download This Paper](#)[Open PDF in Browser](#)Share:     

## Measuring and Explaining Political Sophistication Through Textual Complexity



42 Pages • Posted: 1 Nov 2017

**Kenneth Benoit**

London School of Economics &amp; Political Science (LSE); Trinity College Dublin

**Kevin Munger**

New York University (NYU)

**Arthur Spirling**

New York University

Date Written: October 30, 2017

### Abstract

The sophistication of political communication has been measured using "readability" scores developed from other contexts, but their application out of domain is problematic. We systematically review the shortcomings of previous measures, before developing a new approach, with software, better suited to the task. We use the crowd to perform thousands of pairwise comparisons of text snippets and incorporate these results into a statistical model of comprehension. We include previously excluded features such as parts of speech, and a

# Paper and Software

[Download This Paper](#)[Open PDF in Browser](#)

Share:

## Measuring and Explaining Political Sophistication Through Textual Complexity

42 Pages • Posted: 1 Nov 2017

**Kenneth Benoit**

London School of Economics &amp; Political Science (LSE); Trinity College Dublin

**Kevin Munger**

New York University (NYU)

**Arthur Spirling**

New York University

Date Written: October 30, 2017

### Abstract

The sophistication of political communication has been measured using "readability" scores developed from other contexts, but their application out of domain is problematic. We systematically review the shortcomings of previous measures, before developing a new approach, with software, better suited to the task. We use the crowd to perform thousands of pairwise comparisons of text snippets and incorporate these results into a statistical model of comprehension. We include previously excluded features such as parts of speech, and a



not published build pending build passing coverage 29%

### Code for use in measuring the sophistication of political text

"Measuring and Explaining Political Sophistication Through Textual Complexity" by Kenneth Benoit, Kevin Munger, and Arthur Spirling. This package is built on [quanteda](#).

#### How to install

Using the devtools package:

```
devtools::install_github("kbenoit/sophistication")
```

#### Included Data

new name	original name	description
<code>data_corpus_fifthgrade</code>	<code>fifthCorpus</code>	Fifth-grade reading texts
<code>data_corpus_crimson</code>	<code>crimsonCorpus</code>	Editorials from the Harvard Crimson
<code>data_corpus_partybroadcast</code>	<code>partybroadcastCorpus</code>	UK political party broadcasts
<code>data_corpus_presdebates</code>	<code>presDebateCorpus</code>	US presidential debates 2016

#### How to use

# Paper and Software

[Download This Paper](#)[Open PDF in Browser](#)Share: [f](#) [t](#) [e](#) [m](#) [d](#) [p](#)

## Measuring and Explaining Political Sophistication Through Textual Complexity

42 Pages • Posted: 1 Nov 2017

**Kenneth Benoit**

London School of Economics &amp; Political Science (LSE); Trinity College Dublin

**Kevin Munger**

New York University (NYU)

**Arthur Spirling**

New York University

Date Written: October 30, 2017

### Abstract

The sophistication of political communication has been measured using "readability" scores developed from other contexts, but their application out of domain is problematic. We systematically review the shortcomings of previous measures, before developing a new approach, with software, better suited to the task. We use the crowd to perform thousands of pairwise comparisons of text snippets and incorporate these results into a statistical model of comprehension. We include previously excluded features such as parts of speech, and a

[CRAN](#) [not published](#) [build](#) [personality](#) [build](#) [downing](#) [coverage](#) 29%

### Code for use in measuring the sophistication of political text

"Measuring and Explaining Political Sophistication Through Textual Complexity" by Kenneth Benoit, Kevin Munger, and Arthur Spirling. This package is built on [quanteda](#).

#### How to install

Using the devtools package:

```
devtools::install_github("kbenoit/sophistication")
```

#### Included Data

new name	original name	description
<code>data_corpus_fifthgrade</code>	<code>fifthCorpus</code>	Fifth-grade reading texts
<code>data_corpus_crimson</code>	<code>crimsonCorpus</code>	Editorials from the Harvard Crimson
<code>data_corpus_partybroadcast</code>	<code>partybroadcastCorpus</code>	UK political party broadcasts
<code>data_corpus_presdebates</code>	<code>presDebateCorpus</code>	US presidential debates 2016

#### How to use

[github.com/kbenoit/sophistication](https://github.com/kbenoit/sophistication)

## 4. Style

# Mystery of *The Federalist Papers*

# Mystery of *The Federalist Papers*



# Mystery of *The Federalist Papers*



85 essays published [anonymously](#) in 1787 and 1788

## Mystery of *The Federalist Papers*



85 essays published [anonymously](#) in 1787 and 1788

Generally agreed that Alexander Hamilton wrote 51 essays, John Jay wrote 5 essays, James Madison wrote 14 essays, and 3 essays were written jointly by Hamilton and Madison.

## Mystery of *The Federalist Papers*



85 essays published **anonymously** in 1787 and 1788

Generally agreed that Alexander Hamilton wrote 51 essays, John Jay wrote 5 essays, James Madison wrote 14 essays, and 3 essays were written jointly by Hamilton and Madison.

That leaves 12 that are **disputed**.

## Mystery of *The Federalist Papers*



85 essays published **anonymously** in 1787 and 1788

Generally agreed that Alexander Hamilton wrote 51 essays, John Jay wrote 5 essays, James Madison wrote 14 essays, and 3 essays were written jointly by Hamilton and Madison.

That leaves 12 that are **disputed**.

# Mosteller and Wallace, 1963/4

# Mosteller and Wallace, 1963/4

In essence, they. . .

# Mosteller and Wallace, 1963/4

In essence, they...

Count word frequencies of **function** words (by, from, to, etc.) in the 73 essays with **undisputed** authorship

# Mosteller and Wallace, 1963/4

In essence, they...

Count word frequencies of **function** words (by, from, to, etc.) in the  
73 essays with **undisputed** authorship

then collapse on author to get word frequencies specific to the authors

# Mosteller and Wallace, 1963/4

In essence, they...

Count word frequencies of **function** words (by, from, to, etc.) in the 73 essays with **undisputed** authorship

then collapse on author to get word frequencies specific to the authors

now model these **author-specific rates** with Poisson and negative binomial distributions

# Mosteller and Wallace, 1963/4

In essence, they...

Count word frequencies of **function** words (by, from, to, etc.) in the 73 essays with **undisputed** authorship

then collapse on author to get word frequencies specific to the authors

now model these **author-specific rates** with Poisson and negative binomial distributions

use **Bayes' theorem** to determine the posterior probability that Hamilton (Madison) wrote a particular disputed essay for all such essays

# Mosteller and Wallace, 1963/4

In essence, they...

Count word frequencies of **function** words (by, from, to, etc.) in the 73 essays with **undisputed** authorship

then collapse on author to get word frequencies specific to the authors

now model these **author-specific rates** with Poisson and negative binomial distributions

use **Bayes' theorem** to determine the posterior probability that Hamilton (Madison) wrote a particular disputed essay for all such essays

i.e. they ask "if rates of function word usage are **constant within authors** for these documents, which author was most likely to have written essay  $x$  given the observed function word usage of these authors on the other documents?"

## More Details

# More Details

a	been	had	its
one	the	were	all
but	has	may	only
their	what	also	by
have	more	or	then
when	an	can	her
must	our	there	which
and	do	his	my
things	who	any	down
if	no	so	this
are	even	in	not
some	to	with	as
every	into	now	such
up	would	at	for
is	of	than	upon
your	be	from	it
on	that	was	will
should			

# More Details

may think that sentence length distinguishes authors

a	been	had	its
one	the	were	all
but	has	may	only
their	what	also	by
have	more	or	then
when	an	can	her
must	our	there	which
and	do	his	my
things	who	any	down
if	no	so	this
are	even	in	not
some	to	with	as
every	into	now	such
up	would	at	for
is	of	than	upon
your	be	from	it
on	that	was	will
should			

## More Details

may think that sentence length distinguishes authors, but Hamilton and Madison “practically twins” on this.

a	been	had	its
one	the	were	all
but	has	may	only
their	what	also	by
have	more	or	then
when	an	can	her
must	our	there	which
and	do	his	my
things	who	any	down
if	no	so	this
are	even	in	not
some	to	with	as
every	into	now	such
up	would	at	for
is	of	than	upon
your	be	from	it
on	that	was	will
should			

# More Details

may think that sentence length distinguishes authors, but Hamilton and Madison “practically twins” on this.

a	been	had	its
one	the	were	all
but	has	may	only
their	what	also	by
have	more	or	then
when	an	can	her
must	our	there	which
and	do	his	my
things	who	any	down
if	no	so	this
are	even	in	not
some	to	with	as
every	into	now	such
up	would	at	for
is	of	than	upon
your	be	from	it
on	that	was	will
should			

use function words—conjunctions, prepositions, pronouns—

## More Details

may think that sentence length distinguishes authors, but Hamilton and Madison “practically twins” on this.

a	been	had	its
one	the	were	all
but	has	may	only
their	what	also	by
have	more	or	then
when	an	can	her
must	our	there	which
and	do	his	my
things	who	any	down
if	no	so	this
are	even	in	not
some	to	with	as
every	into	now	such
up	would	at	for
is	of	than	upon
your	be	from	it
on	that	was	will
should			

use function words—conjunctions, prepositions, pronouns—for two (related) reasons:

## More Details

may think that sentence length distinguishes authors, but Hamilton and Madison “practically twins” on this.

a	been	had	its
one	the	were	all
but	has	may	only
their	what	also	by
have	more	or	then
when	an	can	her
must	our	there	which
and	do	his	my
things	who	any	down
if	no	so	this
are	even	in	not
some	to	with	as
every	into	now	such
up	would	at	for
is	of	than	upon
your	be	from	it
on	that	was	will
should			

use function words—conjunctions, prepositions, pronouns—for two (related) reasons:

- ① authors use them unconsciously

## More Details

may think that sentence length distinguishes authors, but Hamilton and Madison “practically twins” on this.

a	been	had	its
one	the	were	all
but	has	may	only
their	what	also	by
have	more	or	then
when	an	can	her
must	our	there	which
and	do	his	my
things	who	any	down
if	no	so	this
are	even	in	not
some	to	with	as
every	into	now	such
up	would	at	for
is	of	than	upon
your	be	from	it
on	that	was	will
should			

use function words—conjunctions, prepositions, pronouns—for two (related) reasons:

- ① authors use them unconsciously
- ② therefore, don't vary much by topic.

## More Details

may think that sentence length distinguishes authors, but Hamilton and Madison “practically twins” on this.

a	been	had	its
one	the	were	all
but	has	may	only
their	what	also	by
have	more	or	then
when	an	can	her
must	our	there	which
and	do	his	my
things	who	any	down
if	no	so	this
are	even	in	not
some	to	with	as
every	into	now	such
up	would	at	for
is	of	than	upon
your	be	from	it
on	that	was	will
should			

use function words—conjunctions, prepositions, pronouns—for two (related) reasons:

- ① authors use them unconsciously
- ② therefore, don't vary much by topic.

NB typically assume one instance of a function word is independent of the next,

## More Details

a	been	had	its
one	the	were	all
but	has	may	only
their	what	also	by
have	more	or	then
when	an	can	her
must	our	there	which
and	do	his	my
things	who	any	down
if	no	so	this
are	even	in	not
some	to	with	as
every	into	now	such
up	would	at	for
is	of	than	upon
your	be	from	it
on	that	was	will
should			

may think that sentence length distinguishes authors, but Hamilton and Madison “practically twins” on this.

use function words—conjunctions, prepositions, pronouns—for two (related) reasons:

- ① authors use them unconsciously
- ② therefore, don't vary much by topic.

NB typically assume one instance of a function word is independent of the next, and use is fixed over a lifetime (and constant within a given text).

## More Details

a	been	had	its
one	the	were	all
but	has	may	only
their	what	also	by
have	more	or	then
when	an	can	her
must	our	there	which
and	do	his	my
things	who	any	down
if	no	so	this
are	even	in	not
some	to	with	as
every	into	now	such
up	would	at	for
is	of	than	upon
your	be	from	it
on	that	was	will
should			

may think that sentence length distinguishes authors, but Hamilton and Madison “practically twins” on this.

use function words—conjunctions, prepositions, pronouns—for two (related) reasons:

- ① authors use them unconsciously
- ② therefore, don't vary much by topic.

NB typically assume one instance of a function word is independent of the next, and use is fixed over a lifetime (and constant within a given text).

→ wrong,

## More Details

a	been	had	its
one	the	were	all
but	has	may	only
their	what	also	by
have	more	or	then
when	an	can	her
must	our	there	which
and	do	his	my
things	who	any	down
if	no	so	this
are	even	in	not
some	to	with	as
every	into	now	such
up	would	at	for
is	of	than	upon
your	be	from	it
on	that	was	will
should			

may think that sentence length distinguishes authors, but Hamilton and Madison “practically twins” on this.

use function words—conjunctions, prepositions, pronouns—for two (related) reasons:

- ① authors use them **unconsciously**
- ② therefore, don't vary much by **topic**.

NB typically assume one instance of a function word is **independent** of the next, and use is fixed over a **lifetime** (and constant within a given text).

→ wrong, but models relying on these assns discriminate well (see Peng & Hengartner on e.g. Austin v Shakespeare)

# The Model (Airoldi et al, 2007)

## The Model (Airoldi et al, 2007)

let  $X_w$  be number of times word  $w$  appears in a document.

## The Model (Airoldi et al, 2007)

let  $X_w$  be number of times word  $w$  appears in a document. Expected rate of occurrence is  $\Theta_w = \omega\mu$

## The Model (Airoldi et al, 2007)

let  $X_w$  be number of times word  $w$  appears in a document. Expected rate of occurrence is  $\Theta_w = \omega\mu$  where  $\mu$  is expected rate of occurrence in a 'reference length' document (say 1000 words)

## The Model (Airoldi et al, 2007)

let  $X_w$  be number of times word  $w$  appears in a document. Expected rate of occurrence is  $\Theta_w = \omega\mu$  where  $\mu$  is expected rate of occurrence in a 'reference length' document (say 1000 words) and  $\omega$  is length of given document as multiple of reference document.

## The Model (Airoldi et al, 2007)

let  $X_w$  be number of times word  $w$  appears in a document. Expected rate of occurrence is  $\Theta_w = \omega\mu$  where  $\mu$  is expected rate of occurrence in a 'reference length' document (say 1000 words) and  $\omega$  is length of given document as multiple of reference document.

so Poisson:

$$P(X_w = x | \Theta_w = (\omega, \mu)) = \frac{e^{\omega\mu}(\omega\mu)^x}{x!}$$

## The Model (Airoldi et al, 2007)

let  $X_w$  be number of times word  $w$  appears in a document. Expected rate of occurrence is  $\Theta_w = \omega\mu$  where  $\mu$  is expected rate of occurrence in a 'reference length' document (say 1000 words) and  $\omega$  is length of given document as multiple of reference document.

so Poisson:

$$P(X_w = x | \Theta_w = (\omega, \mu)) = \frac{e^{\omega\mu}(\omega\mu)^x}{x!}$$

and Negative Binomial (which adds a gamma distributed random effect,  $\delta$ ):

$$NB(X_w = x | \Theta_w = (\omega, \mu, \delta)) = \frac{\gamma(x+k)}{x!\gamma(k)} (\omega\delta)^x (1 + \omega\delta)^{-(x+k)}$$

# Estimation and Inference

# Estimation and Inference

1 using non-informative priors for the parameters of the distributions,

## Estimation and Inference

- 1 using non-informative priors for the parameters of the distributions, obtain the posteriors on the parameters for the authors on the texts we know they wrote.

## Estimation and Inference

- 1 using non-informative priors for the parameters of the distributions, obtain the posteriors on the parameters for the authors on the texts we know they wrote. Record the mean of those posteriors:

## Estimation and Inference

- 1 using non-informative priors for the parameters of the distributions, obtain the posteriors on the parameters for the authors on the texts we know they wrote. Record the mean of those posteriors: they become the estimates of the parameters in the next step.

## Estimation and Inference

- 1 using non-informative priors for the parameters of the distributions, obtain the posteriors on the parameters for the authors on the texts we know they wrote. Record the mean of those posteriors: they become the estimates of the parameters in the next step.
- 2 begin with equal odds of authorship,

## Estimation and Inference

- 1 using **non-informative priors** for the parameters of the distributions, obtain the posteriors on the parameters for the authors on the texts we **know** they wrote. Record the **mean** of those posteriors: they become the estimates of the parameters in the next step.
- 2 begin with equal odds of authorship, and take each of the words combined with the parameter estimates to produce **posterior odds of authorship** (i.e. odds updated by the data we observed).

## Estimation and Inference

- 1 using **non-informative priors** for the parameters of the distributions, obtain the posteriors on the parameters for the authors on the texts we **know** they wrote. Record the **mean** of those posteriors: they become the estimates of the parameters in the next step.
- 2 begin with equal odds of authorship, and take each of the words combined with the parameter estimates to produce **posterior odds of authorship** (i.e. odds updated by the data we observed).

**NB** this requires careful (groups of) word selection,

## Estimation and Inference

- 1 using non-informative priors for the parameters of the distributions, obtain the posteriors on the parameters for the authors on the texts we know they wrote. Record the mean of those posteriors: they become the estimates of the parameters in the next step.
- 2 begin with equal odds of authorship, and take each of the words combined with the parameter estimates to produce posterior odds of authorship (i.e. odds updated by the data we observed).

NB this requires careful (groups of) word selection, and some sensitivity checking (allow priors to vary somewhat)

## Estimation and Inference

- 1 using non-informative priors for the parameters of the distributions, obtain the posteriors on the parameters for the authors on the texts we know they wrote. Record the mean of those posteriors: they become the estimates of the parameters in the next step.
- 2 begin with equal odds of authorship, and take each of the words combined with the parameter estimates to produce posterior odds of authorship (i.e. odds updated by the data we observed).

NB this requires careful (groups of) word selection, and some sensitivity checking (allow priors to vary somewhat)

→ evidence for Madison is overwhelming for most of the disputed papers: ~ a million to one!

## Estimation and Inference

- 1 using non-informative priors for the parameters of the distributions, obtain the posteriors on the parameters for the authors on the texts we know they wrote. Record the mean of those posteriors: they become the estimates of the parameters in the next step.
- 2 begin with equal odds of authorship, and take each of the words combined with the parameter estimates to produce posterior odds of authorship (i.e. odds updated by the data we observed).

NB this requires careful (groups of) word selection, and some sensitivity checking (allow priors to vary somewhat)

- evidence for Madison is overwhelming for most of the disputed papers: ~ a million to one!
- + confirmed by many subsequent analyses (via e.g. machine learning)

# 5. Doing Better

Can we do better?

# Can we do better?

i.e. have I, personally, written a paper about this?

# Can we do better?

i.e. have I, personally, written a paper about this?

YES!

## Can we do better?

i.e. have I, personally, written a paper about this?

YES! SHP consider extension of M&W to arbitrarily large number of documents, authors and tokens.

## Can we do better?

i.e. have I, personally, written a paper about this?

YES! SHP consider extension of M&W to arbitrarily large number of documents, authors and tokens.

Intuition: you are ‘interesting’ if we can determine you were the author/speaker of a speech with relative high probability (on average).

## Can we do better?

i.e. have I, personally, written a paper about this?

YES! SHP consider extension of M&W to arbitrarily large number of documents, authors and tokens.

Intuition: you are ‘interesting’ if we can determine you were the author/speaker of a speech with relative high probability (on average). You are ‘boring’ if we can’t.

## Can we do better?

i.e. have I, personally, written a paper about this?

YES! SHP consider extension of M&W to arbitrarily large number of documents, authors and tokens.

Intuition: you are ‘interesting’ if we can determine you were the author/speaker of a speech with relative high probability (on average). You are ‘boring’ if we can’t.

Data is all speeches by backbenchers in UK HoC, 1935–2018

Formally...

## Formally...

Consider posterior log-odds of authorship for speech  $i$  for speaker  $t$  vs  $s$  ( $\sim$  M&W):

$$\sum_{v \in V} x_{iv} \log \left\{ \frac{\Pr(v|t)}{\Pr(v|s)} \right\}$$

## Formally...

Consider posterior log-odds of authorship for speech  $i$  for speaker  $t$  vs  $s$  ( $\sim$  M&W):

$$\sum_{v \in V} x_{iv} \log \left\{ \frac{\Pr(v|t)}{\Pr(v|s)} \right\}$$

where  $x_{iv}$  is the incidence of token  $v$  in speech  $i$ .

## Formally...

Consider posterior log-odds of authorship for speech  $i$  for speaker  $t$  vs  $s$  ( $\sim$  M&W):

$$\sum_{v \in V} x_{iv} \log \left\{ \frac{\Pr(v|t)}{\Pr(v|s)} \right\}$$

where  $x_{iv}$  is the incidence of token  $v$  in speech  $i$ .

Then, think about average log-odds per token (let  $n_i$  be tokens in speech  $i$ )

$$\sum_{v \in V} (x_{iv}/n_i) \log \left\{ \frac{\Pr(v|t)}{\Pr(v|s)} \right\}$$

## Formally...

Consider posterior log-odds of authorship for speech  $i$  for speaker  $t$  vs  $s$  ( $\sim$  M&W):

$$\sum_{v \in V} x_{iv} \log \left\{ \frac{\Pr(v|t)}{\Pr(v|s)} \right\}$$

where  $x_{iv}$  is the incidence of token  $v$  in speech  $i$ .

Then, think about average log-odds per token (let  $n_i$  be tokens in speech  $i$ )

$$\sum_{v \in V} (x_{iv}/n_i) \log \left\{ \frac{\Pr(v|t)}{\Pr(v|s)} \right\}$$

Then, average over all speakers and speeches (by  $t$ ).

## Formally...

Consider posterior log-odds of authorship for speech  $i$  for speaker  $t$  vs  $s$  ( $\sim$  M&W):

$$\sum_{v \in V} x_{iv} \log \left\{ \frac{\Pr(v|t)}{\Pr(v|s)} \right\}$$

where  $x_{iv}$  is the incidence of token  $v$  in speech  $i$ .

Then, think about average log-odds per token (let  $n_i$  be tokens in speech  $i$ )

$$\sum_{v \in V} (x_{iv}/n_i) \log \left\{ \frac{\Pr(v|t)}{\Pr(v|s)} \right\}$$

Then, average over all speakers and speeches (by  $t$ ).

Variance has closed form analytical expression.

## Formally...

Consider posterior log-odds of authorship for speech  $i$  for speaker  $t$  vs  $s$  ( $\sim$  M&W):

$$\sum_{v \in V} x_{iv} \log \left\{ \frac{\Pr(v|t)}{\Pr(v|s)} \right\}$$

where  $x_{iv}$  is the incidence of token  $v$  in speech  $i$ .

Then, think about average log-odds per token (let  $n_i$  be tokens in speech  $i$ )

$$\sum_{v \in V} (x_{iv}/n_i) \log \left\{ \frac{\Pr(v|t)}{\Pr(v|s)} \right\}$$

Then, average over all speakers and speeches (by  $t$ ).

Variance has closed form analytical expression.

Estimation/fitting generally fast.

# Introducing stylest

# Introducing stylest



[stylest](#), an [R](#) package for

build passing build passing

## stylest (alpha version)

`stylest` estimates speaker style distinctiveness.

### Installation

You can install `stylest` from github with:

# Introducing stylest



[stylest](#), an [R](#) package for

- ▶ estimating distinctiveness of speakers/authors/anything

[build](#) passing [build](#) passing

## stylest (alpha version)

`stylest` estimates speaker style distinctiveness.

### Installation

You can install `stylest` from github with:

# Introducing stylest



[stylest](#), an [R](#) package for

- ▶ estimating distinctiveness of speakers/authors/anything
- ▶ estimating [similarity](#) of authors/documents to each other

build passing build passing

## stylest (alpha version)

`stylest` estimates speaker style distinctiveness.

### Installation

You can install `stylest` from github with:

# Introducing stylest



[stylest](#), an [R](#) package for

- ▶ estimating distinctiveness of speakers/authors/anything
- ▶ estimating [similarity](#) of authors/documents to each other
- ▶ identifying most likely author of “[mystery](#)” texts

build passing  build passing 

## stylest (alpha version)

`stylest` estimates speaker style distinctiveness.

### Installation

You can install `stylest` from github with:

# Introducing stylest



[stylest](#), an [R](#) package for

- ▶ estimating distinctiveness of speakers/authors/anything
- ▶ estimating [similarity](#) of authors/documents to each other
- ▶ identifying most likely author of “[mystery](#)” texts

Install from:

build passing build passing

## stylest (alpha version)

`stylest` estimates speaker style distinctiveness.

### Installation

You can install `stylest` from github with:

# Introducing stylest



[stylest](#), an R package for

- ▶ estimating distinctiveness of speakers/authors/anything
- ▶ estimating [similarity](#) of authors/documents to each other
- ▶ identifying most likely author of “[mystery](#)” texts

Install from:

<https://github.com/leslie-huang/stylest>

The screenshot shows a GitHub repository page for 'stylest (alpha version)'. At the top, there are two green 'build passing' status indicators. Below them, the repository name 'stylest (alpha version)' is displayed in bold. A brief description follows: 'stylest estimates speaker style distinctiveness.' Under the heading 'Installation', it says 'You can install stylest from github with:'.

# Validation, $\mathbb{D}_t$

## Validation, $\mathbb{D}_t$

Look at 'top 20' (most interesting) vs 'bottom 20' (most boring) backbenchers,

## Validation, $\mathbb{D}_t$

Look at 'top 20' (most interesting) vs 'bottom 20' (most boring) backbenchers, either side of Blair 1997 landslide.

## Validation, $\mathbb{D}_t$

Look at 'top 20' (most interesting) vs 'bottom 20' (most boring) backbenchers, either side of Blair 1997 landslide.

Compare mentions in *The Times* (for relevant parliamentary session), non-parametric means test ( $p < 0.05$ )

## Validation, $\mathbb{D}_t$

Look at 'top 20' (most interesting) vs 'bottom 20' (most boring) backbenchers, either side of Blair 1997 landslide.

Compare mentions in *The Times* (for relevant parliamentary session),  
non-parametric means test ( $p < 0.05$ )



## Validation, $\mathbb{D}_t$

Look at 'top 20' (most interesting) vs 'bottom 20' (most boring) backbenchers, either side of Blair 1997 landslide.

Compare mentions in *The Times* (for relevant parliamentary session),  
non-parametric means test ( $p < 0.05$ )



# Validation, $\mathbb{D}_t$

Look at 'top 20' (most interesting) vs 'bottom 20' (most boring) backbenchers, either side of Blair 1997 landslide.

Compare mentions in *The Times* (for relevant parliamentary session), non-parametric means test ( $p < 0.05$ )



# Identifying ‘intruders’: James Joyce

# Identifying ‘intruders’: James Joyce

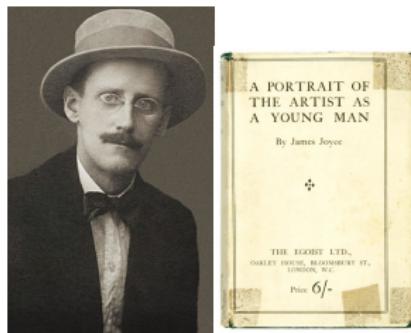
If model works,

# Identifying ‘intruders’: James Joyce

If model works, it should give  
‘intruders’ high distinctiveness.

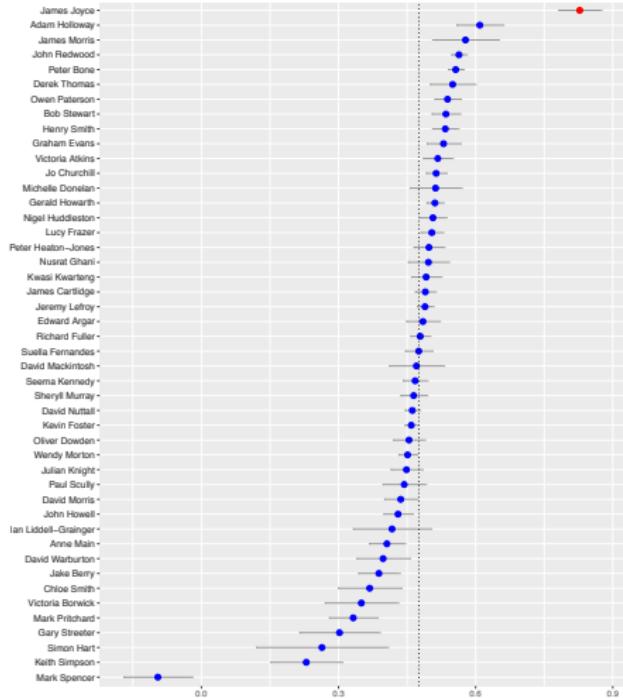
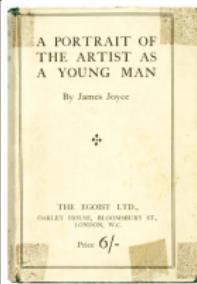
# Identifying ‘intruders’: James Joyce

If model works, it should give  
‘intruders’ high distinctiveness.



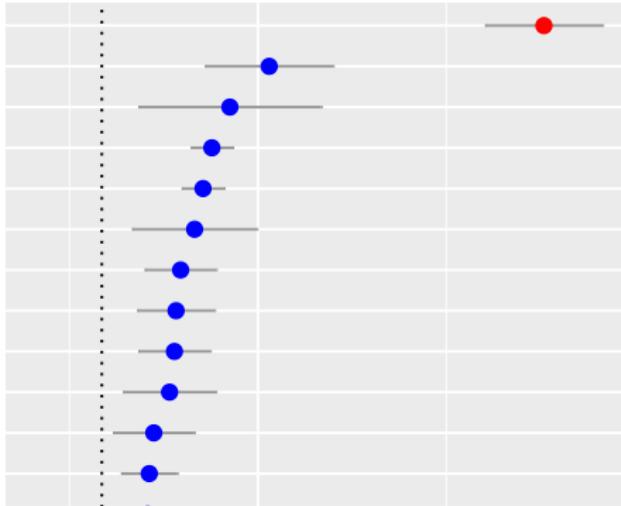
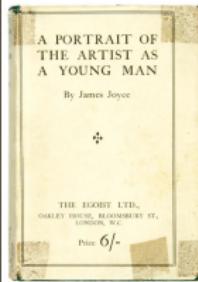
# Identifying ‘intruders’: James Joyce

If model works, it should give  
‘intruders’ high distinctiveness.



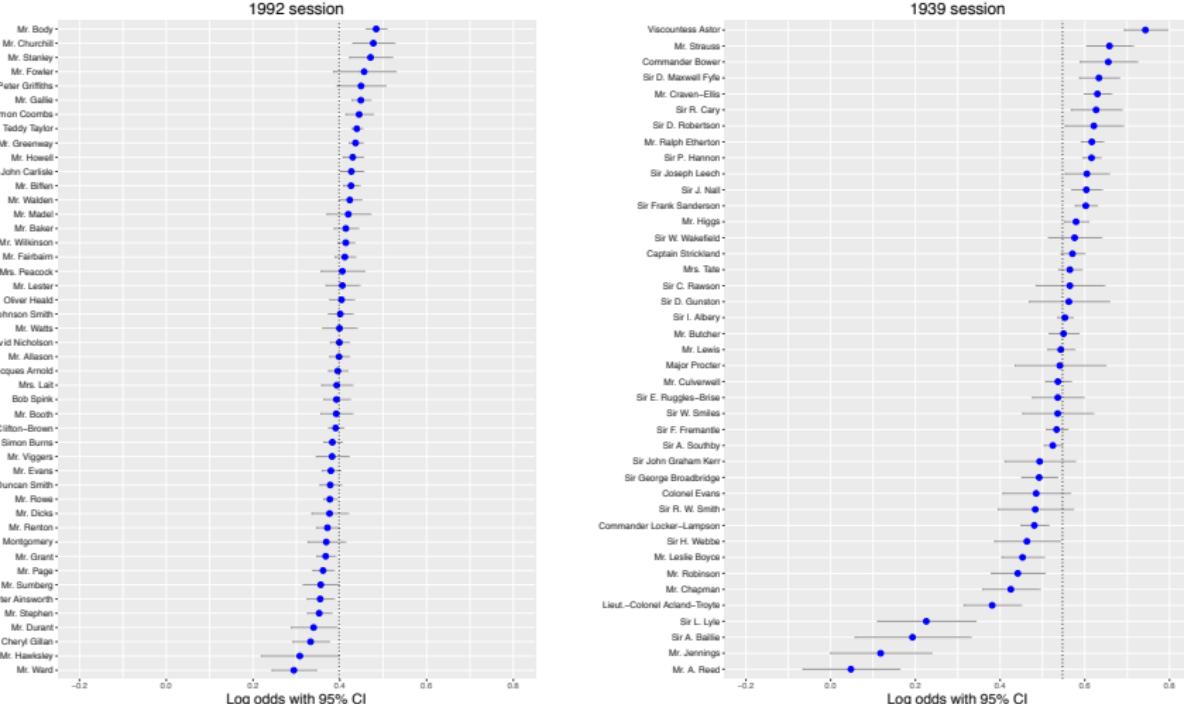
# Identifying 'intruders': James Joyce

If model works, it should give  
'intruders' high distinctiveness.



# Comparing Sessions: 1992 v 1939

# Comparing Sessions: 1992 v 1939



# Words

0

# Words

Influential words are those used often by some speakers relative to other speakers, weighted by how frequently the word is used in practice.

# Words

**Influential** words are those used often by some speakers relative to other speakers, weighted by how frequently the word is used in practice.

e.g. If a speaker uses 'europe' three times as often as other speakers, *and* she uses that term a great deal, it will be influential.

# Words

**Influential** words are those used often by some speakers relative to other speakers, weighted by how frequently the word is used in practice.

e.g. If a speaker uses 'europe' three times as often as other speakers, *and* she uses that term a great deal, it will be influential.

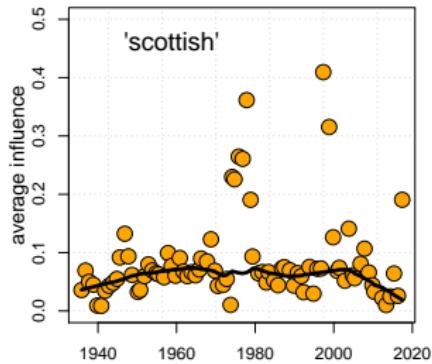
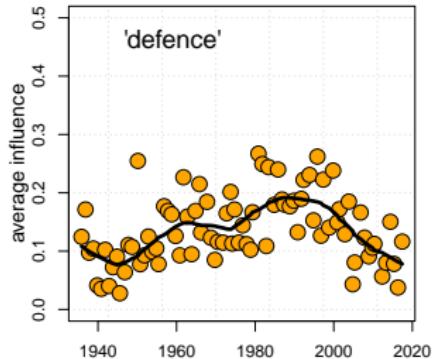
→ tend to see a lot of **stop words** as influential.

# Words

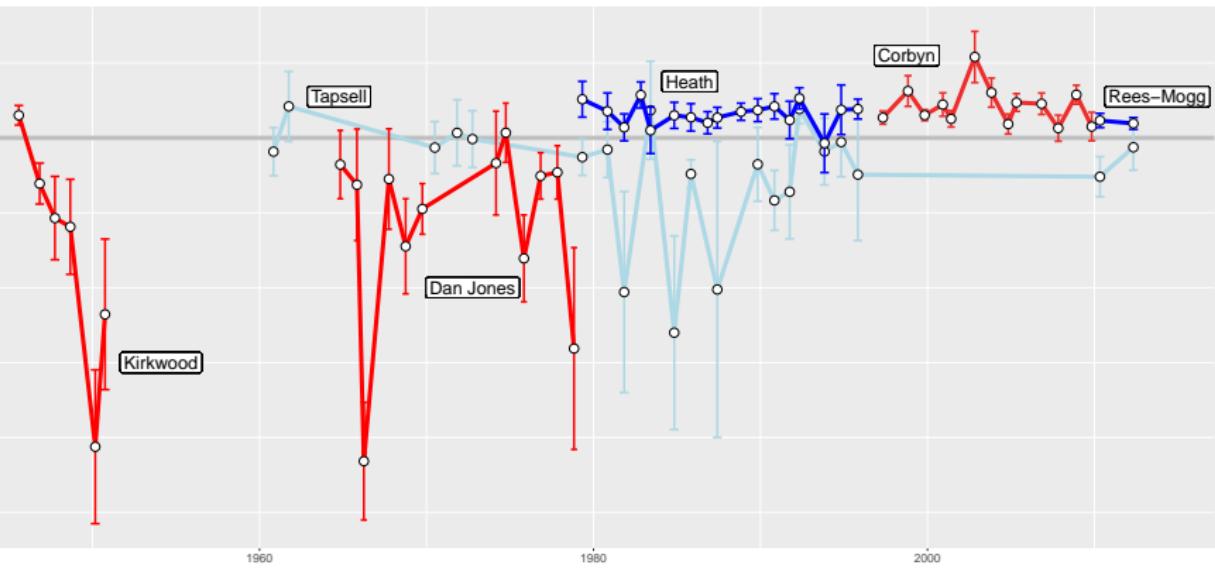
Influential words are those used often by some speakers relative to other speakers, weighted by how frequently the word is used in practice.

e.g. If a speaker uses 'europe' three times as often as other speakers, *and* she uses that term a great deal, it will be influential.

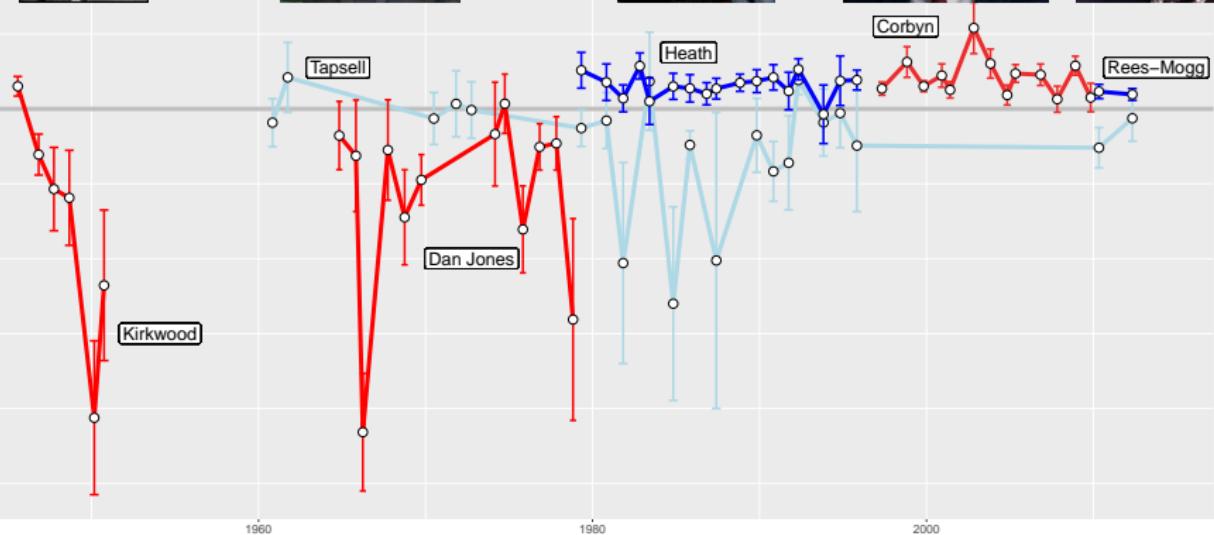
→ tend to see a lot of stop words as influential.



# Usual Suspects (Z normalization)

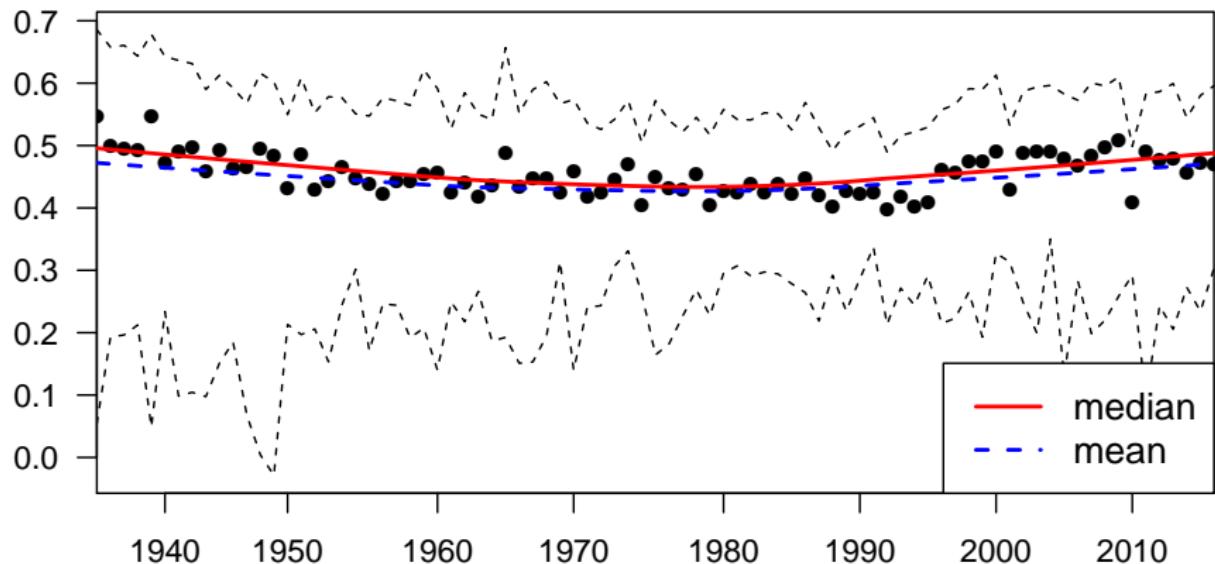


# Usual Suspects ( $Z$ normalization)



# Average Level of Boringness is Constant!

# Average Level of Boringness is Constant!



# Not much happening with distribution!

# Not much happening with distribution!

# Interestingness is decreasing in seniority

# Interestingness is decreasing in seniority

	(1)	(2)	(3)
experience	—	—	—
demoted	+	—	+
MP-fixed effects	✗	✓	✓
Session-fixed effects	✗	✗	✓

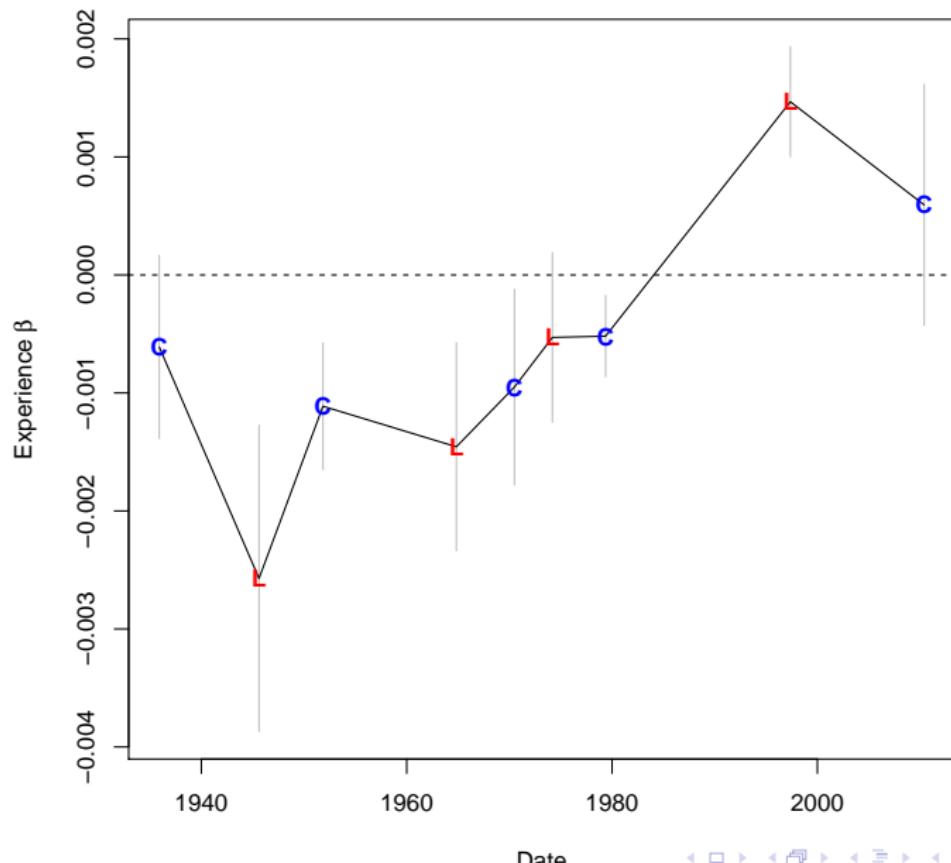
# Interestingness is decreasing in seniority

	(1)	(2)	(3)
experience	—	—	—
demoted	+	—	+
MP-fixed effects	✗	✓	✓
Session-fixed effects	✗	✗	✓

Dependent variable: 'distinctiveness' (in average log-odds terms)

# Effect of Seniority is Changing

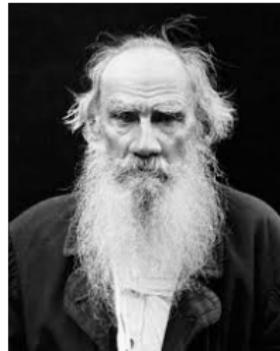
# Effect of Seniority is Changing



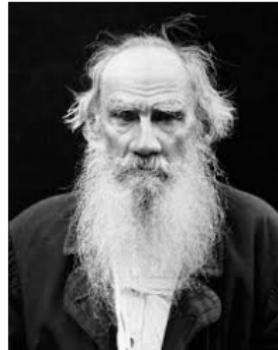
# **6. Statistical Properties**

# Partner Exercise

# Partner Exercise



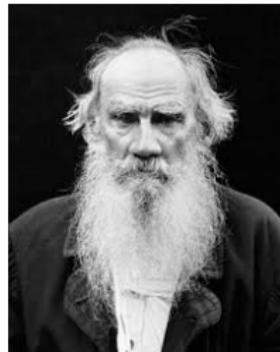
## Partner Exercise



Suppose you wanted to compare the novels of Leo Tolstoy and J.D. Salinger.



## Partner Exercise

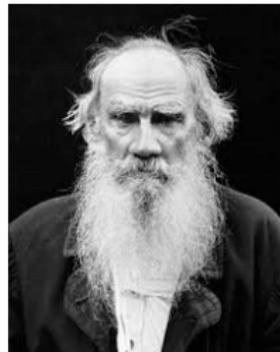


Suppose you wanted to compare the novels of Leo Tolstoy and J.D. Salinger.



- 1 You estimate the FRE score for *War and Peace* and compare it to the FRE of *The Catcher in the Rye*. Which estimate are you more confident in?

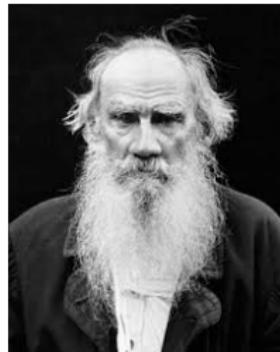
## Partner Exercise



Suppose you wanted to compare the novels of Leo Tolstoy and J.D. Salinger.

- 1 You estimate the FRE score for *War and Peace* and compare it to the FRE of *The Catcher in the Rye*. Which estimate are you more confident in? Why?

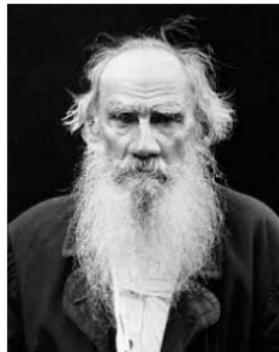
## Partner Exercise



Suppose you wanted to compare the novels of Leo Tolstoy and J.D. Salinger.

- 1 You estimate the FRE score for *War and Peace* and compare it to the FRE of *The Catcher in the Rye*. Which estimate are you more confident in? Why?
- 2 You estimate the (average) FRE scores for all their novels, respectively. Which estimate (for Tolstoy or Salinger) are you more confident in?

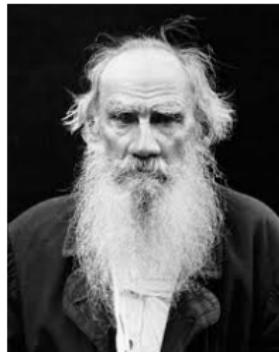
## Partner Exercise



Suppose you wanted to compare the novels of Leo Tolstoy and J.D. Salinger.

- 1 You estimate the FRE score for *War and Peace* and compare it to the FRE of *The Catcher in the Rye*. Which estimate are you more confident in? Why?
- 2 You estimate the (average) FRE scores for all their novels, respectively. Which estimate (for Tolstoy or Salinger) are you more confident in? Why?

## Partner Exercise



Suppose you wanted to compare the novels of Leo Tolstoy and J.D. Salinger.

- 1 You estimate the FRE score for *War and Peace* and compare it to the FRE of *The Catcher in the Rye*. Which estimate are you more confident in? Why?
- 2 You estimate the (average) FRE scores for all their novels, respectively. Which estimate (for Tolstoy or Salinger) are you more confident in? Why?

# Sampling and Uncertainty

# Sampling and Uncertainty

To now,

# Sampling and Uncertainty

To now, we've been concerned with point estimates

# Sampling and Uncertainty

To now, we've been concerned with **point estimates**

e.g. the lexical diversity of a story is 0.43,

# Sampling and Uncertainty

To now, we've been concerned with **point estimates**

e.g. the lexical diversity of a story is 0.43, the FRE of a speech is 55.2 etc

# Sampling and Uncertainty

To now, we've been concerned with **point estimates**

e.g. the lexical diversity of a story is 0.43, the FRE of a speech is 55.2 etc

But this is **unsatisfying**:

# Sampling and Uncertainty

To now, we've been concerned with **point estimates**

e.g. the lexical diversity of a story is 0.43, the FRE of a speech is 55.2 etc

But this is **unsatisfying**: it says nothing of the **variance** of such estimates,

# Sampling and Uncertainty

To now, we've been concerned with **point estimates**

e.g. the lexical diversity of a story is 0.43, the FRE of a speech is 55.2 etc

But this is **unsatisfying**: it says nothing of the **variance** of such estimates, yet surely we are **more certain** of an estimate from a **long** text (or many texts) than we are from a **short** text.

# Sampling and Uncertainty

To now, we've been concerned with **point estimates**

e.g. the lexical diversity of a story is 0.43, the FRE of a speech is 55.2 etc

But this is **unsatisfying**: it says nothing of the **variance** of such estimates, yet surely we are **more certain** of an estimate from a **long** text (or many texts) than we are from a **short** text.

e.g. how much would you trust a one word response vs a 1000-word speech from a member of parliament in terms of its complexity or policy content?

# Sampling and Uncertainty

To now, we've been concerned with **point estimates**

e.g. the lexical diversity of a story is 0.43, the FRE of a speech is 55.2 etc

But this is **unsatisfying**: it says nothing of the **variance** of such estimates, yet surely we are **more certain** of an estimate from a **long** text (or many texts) than we are from a **short** text.

e.g. how much would you trust a one word response vs a 1000-word speech from a member of parliament in terms of its complexity or policy content?

→ think a little more systematically about the **sampling distribution** of a statistic.

# Sampling Distributions: Reminder

## Sampling Distributions: Reminder

Suppose we are interested in the population mean,  $\mu$  and we our we use the sample mean  $\bar{x}$

## Sampling Distributions: Reminder

Suppose we are interested in the population mean,  $\mu$  and we use the sample mean  $\bar{x}$  as our estimator of it.

## Sampling Distributions: Reminder

Suppose we are interested in the population mean,  $\mu$  and we use the sample mean  $\bar{x}$  as our estimator of it.

We want the sampling distribution of sample mean,

## Sampling Distributions: Reminder

Suppose we are interested in the population mean,  $\mu$  and we use the sample mean  $\bar{x}$  as our estimator of it.

We want the sampling distribution of sample mean, i.e. the distribution of the sample mean for all possible samples we *could have* drawn.

## Sampling Distributions: Reminder

Suppose we are interested in the population mean,  $\mu$  and we use the sample mean  $\bar{x}$  as our estimator of it.

We want the sampling distribution of sample mean, i.e. the distribution of the sample mean for all possible samples we *could have* drawn.

→ important,

## Sampling Distributions: Reminder

Suppose we are interested in the population mean,  $\mu$  and we use the sample mean  $\bar{x}$  as our estimator of it.

We want the sampling distribution of sample mean, i.e. the distribution of the sample mean for all possible samples we *could have* drawn.

→ important, because it tells us the uncertainty around our statistic,

## Sampling Distributions: Reminder

Suppose we are interested in the population mean,  $\mu$  and we use the sample mean  $\bar{x}$  as our estimator of it.

We want the sampling distribution of sample mean, i.e. the distribution of the sample mean for all possible samples we *could have* drawn.

- important, because it tells us the uncertainty around our statistic, and we can use that to produce

## Sampling Distributions: Reminder

Suppose we are interested in the population mean,  $\mu$  and we use the sample mean  $\bar{x}$  as our estimator of it.

We want the sampling distribution of sample mean, i.e. the distribution of the sample mean for all possible samples we *could have* drawn.

- important, because it tells us the uncertainty around our statistic, and we can use that to produce e.g. confidence intervals

## Sampling Distributions: Reminder

Suppose we are interested in the population mean,  $\mu$  and we use the sample mean  $\bar{x}$  as our estimator of it.

We want the sampling distribution of sample mean, i.e. the distribution of the sample mean for all possible samples we *could have* drawn.

- important, because it tells us the uncertainty around our statistic, and we can use that to produce e.g. confidence intervals and make statements about the statistical significance of differences between means of different groups.

## Sampling Distributions: Reminder

Suppose we are interested in the population mean,  $\mu$  and we use the sample mean  $\bar{x}$  as our estimator of it.

We want the sampling distribution of sample mean, i.e. the distribution of the sample mean for all possible samples we *could have* drawn.

- important, because it tells us the uncertainty around our statistic, and we can use that to produce e.g. confidence intervals and make statements about the statistical significance of differences between means of different groups.

# Normal Case

## Normal Case

For a large enough number  
of samples of sufficient  
size,

## Normal Case

For a large enough number of samples of sufficient size, that sampling distribution is **normally distributed** (via the Central Limit Theorem)—

## Normal Case

For a large enough number of samples of sufficient size, that sampling distribution is **normally distributed** (via the Central Limit Theorem)—

$$\bar{x} \sim \left( \mu, \frac{\sigma^2}{n} \right)$$

## Normal Case

For a large enough number of samples of sufficient size, that sampling distribution is **normally distributed** (via the Central Limit Theorem)—  
 $\bar{x} \sim \left( \mu, \frac{\sigma^2}{n} \right)$ .

**NB** We call the standard deviation of the sampling distribution the **standard error** of the statistic.

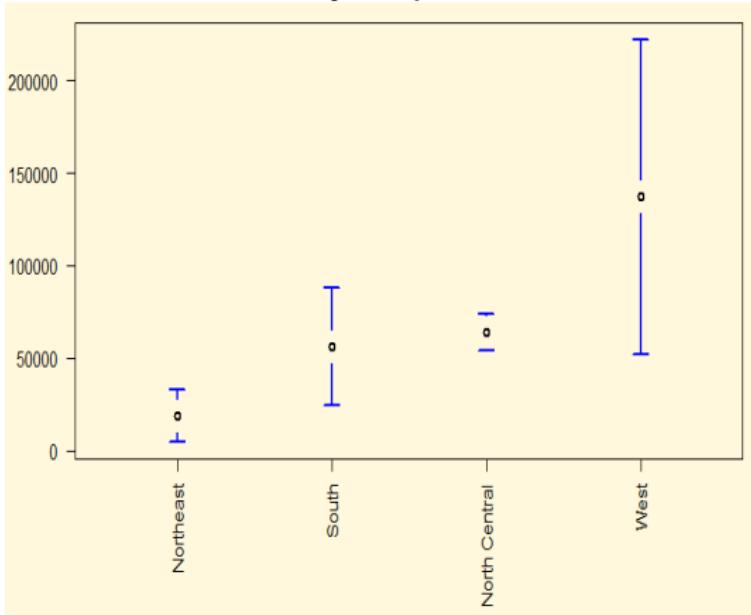
## Normal Case

For a large enough number of samples of sufficient size, that sampling distribution is **normally distributed** (via the Central Limit Theorem)—

$$\bar{x} \sim \left( \mu, \frac{\sigma^2}{n} \right).$$

**NB** We call the standard deviation of the sampling distribution the **standard error** of the statistic.

Very helpful!



# Sampling Distributions for Text

# Sampling Distributions for Text

Need to first think about data generating process by which author intent or characteristic  $\pi$  becomes realized message  $\tau$ .

# Sampling Distributions for Text

Need to first think about data generating process by which author intent or characteristic  $\pi$  becomes realized message  $\tau$ .

This mapping is assumed to be *stochastic* in the sense that it would not be the same 'every time' (even if  $\pi$  were constant).

# Sampling Distributions for Text

Need to first think about data generating process by which author intent or characteristic  $\pi$  becomes realized message  $\tau$ .

This mapping is assumed to be *stochastic* in the sense that it would not be the same 'every time' (even if  $\pi$  were constant).

btw in politics, for strategic reasons,

# Sampling Distributions for Text

Need to first think about data generating process by which author intent or characteristic  $\pi$  becomes realized message  $\tau$ .

This mapping is assumed to be *stochastic* in the sense that it would not be the same 'every time' (even if  $\pi$  were constant).

btw in politics, for strategic reasons,  $\pi$  might not even be the author's true position/complexity/diversity on an issue

# Sampling Distributions for Text

Need to first think about data generating process by which author intent or characteristic  $\pi$  becomes realized message  $\tau$ .

This mapping is assumed to be *stochastic* in the sense that it would not be the same 'every time' (even if  $\pi$  were constant).

btw in politics, for strategic reasons,  $\pi$  might not even be the author's true position/complexity/diversity on an issue

But what is the sampling distribution of FRE for a passage?

# Sampling Distributions for Text

Need to first think about data generating process by which author intent or characteristic  $\pi$  becomes realized message  $\tau$ .

This mapping is assumed to be *stochastic* in the sense that it would not be the same 'every time' (even if  $\pi$  were constant).

btw in politics, for strategic reasons,  $\pi$  might not even be the author's true position/complexity/diversity on an issue

But what is the sampling distribution of FRE for a passage? Is it normal?

# Sampling Distributions for Text

Need to first think about data generating process by which author intent or characteristic  $\pi$  becomes realized message  $\tau$ .

This mapping is assumed to be *stochastic* in the sense that it would not be the same 'every time' (even if  $\pi$  were constant).

btw in politics, for strategic reasons,  $\pi$  might not even be the author's true position/complexity/diversity on an issue

But what is the sampling distribution of FRE for a passage? Is it normal?  
Maybe, maybe not.

# Sampling Distributions for Text

Need to first think about data generating process by which author intent or characteristic  $\pi$  becomes realized message  $\tau$ .

This mapping is assumed to be *stochastic* in the sense that it would not be the same 'every time' (even if  $\pi$  were constant).

btw in politics, for strategic reasons,  $\pi$  might not even be the author's true position/complexity/diversity on an issue

But what is the sampling distribution of FRE for a passage? Is it normal?  
Maybe, maybe not.

→ difficult to know how we should calculate the sampling distribution and thus the standard error.

# Bootstrapping

# Bootstrapping

Bootstrapping is a method to obtain the properties of an estimator

# Bootstrapping

Bootstrapping is a method to obtain the properties of an estimator—the variance, here—via random sampling with replacement from our sample.

# Bootstrapping

Bootstrapping is a method to obtain the properties of an estimator—the variance, here—via random sampling with replacement from our sample.

It treats the sample as the population,

# Bootstrapping

Bootstrapping is a method to obtain the properties of an estimator—the variance, here—via random sampling with replacement from our sample.

It treats the sample as the population, and draws (say, 1000) samples from it,

# Bootstrapping

Bootstrapping is a method to obtain the properties of an estimator—the variance, here—via random sampling with replacement from our sample.

It treats the sample as the population, and draws (say, 1000) samples from it, each time calculating some statistic of interest.

# Bootstrapping

Bootstrapping is a method to obtain the properties of an estimator—the variance, here—via random sampling with replacement from our sample.

It treats the sample as the population, and draws (say, 1000) samples from it, each time calculating some statistic of interest. Putting those statistics together yields a sampling distribution giving us access to the standard error etc.

# Bootstrapping

Bootstrapping is a method to obtain the properties of an estimator—the variance, here—via random sampling with replacement from our sample.

It treats the sample as the population, and draws (say, 1000) samples from it, each time calculating some statistic of interest. Putting those statistics together yields a sampling distribution giving us access to the standard error etc.

NB it is a simulation method,

# Bootstrapping

Bootstrapping is a method to obtain the properties of an estimator—the variance, here—via random sampling with replacement from our sample.

It treats the sample as the population, and draws (say, 1000) samples from it, each time calculating some statistic of interest. Putting those statistics together yields a sampling distribution giving us access to the standard error etc.

NB it is a simulation method, in the sense that we are not analytically deriving the formula for the sampling distribution, but approximating it via random sampling.

# Bootstrapping

Bootstrapping is a method to obtain the properties of an estimator—the variance, here—via random sampling with replacement from our sample.

It treats the sample as the population, and draws (say, 1000) samples from it, each time calculating some statistic of interest. Putting those statistics together yields a sampling distribution giving us access to the standard error etc.

NB it is a simulation method, in the sense that we are not analytically deriving the formula for the sampling distribution, but approximating it via random sampling.

Remarkably,

# Bootstrapping

Bootstrapping is a method to obtain the properties of an estimator—the variance, here—via random sampling with replacement from our sample.

It treats the sample as the population, and draws (say, 1000) samples from it, each time calculating some statistic of interest. Putting those statistics together yields a sampling distribution giving us access to the standard error etc.

NB it is a simulation method, in the sense that we are not analytically deriving the formula for the sampling distribution, but approximating it via random sampling.

Remarkably, it works well,

# Bootstrapping

Bootstrapping is a method to obtain the properties of an estimator—the variance, here—via random sampling with replacement from our sample.

It treats the sample as the population, and draws (say, 1000) samples from it, each time calculating some statistic of interest. Putting those statistics together yields a sampling distribution giving us access to the standard error etc.

NB it is a simulation method, in the sense that we are not analytically deriving the formula for the sampling distribution, but approximating it via random sampling.

Remarkably, it works well, even in quite small samples (e.g.  $N < 20$ )

# Bootstrapping

Bootstrapping is a method to obtain the properties of an estimator—the variance, here—via random sampling with replacement from our sample.

It treats the sample as the population, and draws (say, 1000) samples from it, each time calculating some statistic of interest. Putting those statistics together yields a sampling distribution giving us access to the standard error etc.

NB it is a simulation method, in the sense that we are not analytically deriving the formula for the sampling distribution, but approximating it via random sampling.

Remarkably, it works well, even in quite small samples (e.g.  $N < 20$ )

NB many forms:

# Bootstrapping

Bootstrapping is a method to obtain the properties of an estimator—the variance, here—via random sampling with replacement from our sample.

It treats the sample as the population, and draws (say, 1000) samples from it, each time calculating some statistic of interest. Putting those statistics together yields a sampling distribution giving us access to the standard error etc.

NB it is a simulation method, in the sense that we are not analytically deriving the formula for the sampling distribution, but approximating it via random sampling.

Remarkably, it works well, even in quite small samples (e.g.  $N < 20$ )

NB many forms: non-parametric is most common,

# Bootstrapping

Bootstrapping is a method to obtain the properties of an estimator—the variance, here—via random sampling with replacement from our sample.

It treats the sample as the population, and draws (say, 1000) samples from it, each time calculating some statistic of interest. Putting those statistics together yields a sampling distribution giving us access to the standard error etc.

NB it is a simulation method, in the sense that we are not analytically deriving the formula for the sampling distribution, but approximating it via random sampling.

Remarkably, it works well, even in quite small samples (e.g.  $N < 20$ )

NB many forms: non-parametric is most common, though parametric is more precise (but requires additional assumptions)

# Bootstrap Example

## Bootstrap Example

Have simple linear model,  $n = 20$   
of form  $y_i = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \epsilon_i$

## Bootstrap Example

Have simple linear model,  $n = 20$   
of form  $y_i = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \epsilon_i$

Want to know distribution of  $R^2$ ,

## Bootstrap Example

Have simple linear model,  $n = 20$   
of form  $y_i = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \epsilon_i$

Want to know distribution of  $R^2$ ,  
via bootstrap

## Bootstrap Example

Have simple linear model,  $n = 20$   
of form  $y_i = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \epsilon_i$

Want to know distribution of  $R^2$ ,  
via bootstrap

so resample data ( $n = 20$  every time),

## Bootstrap Example

Have simple linear model,  $n = 20$   
of form  $y_i = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \epsilon_i$

Want to know distribution of  $R^2$ ,  
via bootstrap

so resample data ( $n = 20$  every time),  
and record  $R^2$

## Bootstrap Example

Have simple linear model,  $n = 20$   
of form  $y_i = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \epsilon_i$

Want to know distribution of  $R^2$ ,  
via bootstrap

so resample data ( $n = 20$  every time),  
and record  $R^2$ —then plot...

## Bootstrap Example

Have simple linear model,  $n = 20$   
of form  $y_i = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \epsilon_i$

Want to know distribution of  $R^2$ ,  
via bootstrap

so resample data ( $n = 20$  every time),  
and record  $R^2$ —then plot...

## Bootstrap Example

Have simple linear model,  $n = 20$   
of form  $y_i = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \epsilon_i$

Want to know distribution of  $R^2$ ,  
via bootstrap

so resample data ( $n = 20$  every time),  
and record  $R^2$ —then plot...

## Bootstrap Example

Have simple linear model,  $n = 20$   
of form  $y_i = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \epsilon_i$

Want to know distribution of  $R^2$ ,  
via bootstrap

so resample data ( $n = 20$  every time),  
and record  $R^2$ —then plot...

# Bootstrap Unit

# Bootstrap Unit

When we have a document,

# Bootstrap Unit

When we have a document, need to think about what it represents a sample of.

## Bootstrap Unit

When we have a document, need to think about what it represents a sample of.  
so tokens?

## Bootstrap Unit

When we have a document, need to think about what it represents a sample of.

so tokens? paragraphs?

# Bootstrap Unit

When we have a document, need to think about what it represents a sample of.

so tokens? paragraphs? sentences?

## Bootstrap Unit

When we have a document, need to think about what it represents a sample of.

so tokens? paragraphs? sentences?

Benoit, Laver and Mikhalov (2009) use (quasi-)sentence-level bootstrap,

## Bootstrap Unit

When we have a document, need to think about what it represents a sample of.

so tokens? paragraphs? sentences?

Benoit, Laver and Mikhalev (2009) use (quasi-)sentence-level bootstrap, which makes sense for manifestos:

## Bootstrap Unit

When we have a document, need to think about what it represents a sample of.

so tokens? paragraphs? sentences?

Benoit, Laver and Mikhalev (2009) use (quasi-)sentence-level bootstrap, which makes sense for manifestos: natural unit in which they are written.

## Bootstrap Unit

When we have a document, need to think about what it represents a sample of.

so tokens? paragraphs? sentences?

Benoit, Laver and Mikhalev (2009) use (quasi-)sentence-level bootstrap, which makes sense for manifestos: natural unit in which they are written.

tho this requires segmenting the text into sentences (i.e. cannot use DTM),

## Bootstrap Unit

When we have a document, need to think about what it represents a sample of.

so tokens? paragraphs? sentences?

Benoit, Laver and Mikhalev (2009) use (quasi-)sentence-level bootstrap, which makes sense for manifestos: natural unit in which they are written.

tho this requires segmenting the text into sentences (i.e. cannot use DTM), and performing the relevant operation (e.g. FRE) as many times as there are sentences.

## Bootstrap Unit

When we have a document, need to think about what it represents a sample of.

so tokens? paragraphs? sentences?

Benoit, Laver and Mikhalev (2009) use (quasi-)sentence-level bootstrap, which makes sense for manifestos: natural unit in which they are written.

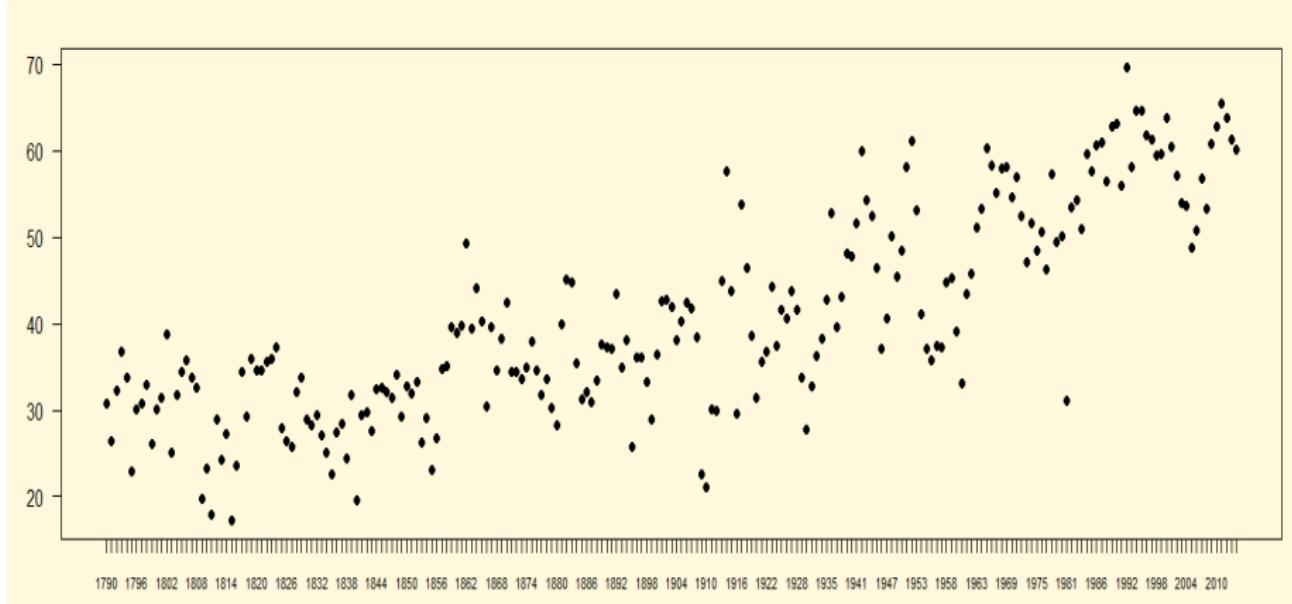
tho this requires segmenting the text into sentences (i.e. cannot use DTM), and performing the relevant operation (e.g. FRE) as many times as there are sentences.

btw long texts give rise to smaller SEs than short ones, which makes sense!

# SOU: 1000 bootstrap samples

0

# SOU: 1000 bootstrap samples



# SOU: 1000 bootstrap samples

